Model Uncertainty and Uncertainty Quantification

Merlise Clyde
Duke University
http://stat.duke.edu/∼clyde

SIAM UQ18 - April 17, 2018

Uncertainty Quantification

Wikipedia:

> *Uncertainty Quantification (UQ) is the science of quantitative characterizing and reduction of uncertainties*
>
> ...

Uncertainty Quantification

Wikipedia:

> *Uncertainty Quantification (UQ) is the science of quantitative characterizing and reduction of uncertainties*
>
> ...

parameter parameters $\boldsymbol{\theta}$ in the model that are unknown

Uncertainty Quantification

Wikipedia:

> *Uncertainty Quantification (UQ) is the science of quantitative characterizing and reduction of uncertainties*
> . . .

parameter	parameters $\boldsymbol{\theta}$ in the model that are unknown
inputs	measurement error in model inputs $\mathbf{x}$

Uncertainty Quantification

Wikipedia:

> *Uncertainty Quantification (UQ) is the science of quantitative characterizing and reduction of uncertainties*
>
> . . .

parameter	parameters $\boldsymbol{\theta}$ in the model that are unknown
inputs	measurement error in model inputs $\mathbf{x}$
algorithmic	induced by approximating model

Uncertainty Quantification

Wikipedia:

> *Uncertainty Quantification (UQ) is the science of quantitative characterizing and reduction of uncertainties* ...

parameter	parameters $\boldsymbol{\theta}$ in the model that are unknown
inputs	measurement error in model inputs $\mathbf{x}$
algorithmic	induced by approximating model
experimental	observation error in response Y at input $\mathbf{x}$

Uncertainty Quantification

Wikipedia:

Uncertainty Quantification (UQ) is the science of quantitative characterizing and reduction of uncertainties ...

parameter	parameters $\boldsymbol{\theta}$ in the model that are unknown
inputs	measurement error in model inputs $\mathbf{x}$
algorithmic	induced by approximating model
experimental	observation error in response Y at input $\mathbf{x}$
model	structural uncertainty about the model/data generating process

Uncertainty Quantification

Wikipedia:

Uncertainty Quantification (UQ) is the science of quantitative characterizing and reduction of uncertainties ...

parameter	parameters $\boldsymbol{\theta}$ in the model that are unknown
inputs	measurement error in model inputs $\mathbf{x}$
algorithmic	induced by approximating model
experimental	observation error in response Y at input $\mathbf{x}$
model	structural uncertainty about the model/data generating process
predictive	interpolation or extrapolation of model at new $\mathbf{x}$

Uncertainty Quantification

Wikipedia:

> *Uncertainty Quantification (UQ) is the science of quantitative characterizing and reduction of uncertainties ...*

parameter	parameters $\boldsymbol{\theta}$ in the model that are unknown
inputs	measurement error in model inputs $\mathbf{x}$
algorithmic	induced by approximating model
experimental	observation error in response Y at input $\mathbf{x}$
model	structural uncertainty about the model/data generating process
predictive	interpolation or extrapolation of model at new $\mathbf{x}$

Predictive uncertainty: reducible error + irreducible error

Uncertainty Quantification

Wikipedia:

> *Uncertainty Quantification (UQ) is the science of quantitative characterizing and reduction of uncertainties …*

parameter	parameters $\boldsymbol{\theta}$ in the model that are unknown
inputs	measurement error in model inputs $\mathbf{x}$
algorithmic	induced by approximating model
experimental	observation error in response Y at input $\mathbf{x}$
model	structural uncertainty about the model/data generating process
predictive	interpolation or extrapolation of model at new $\mathbf{x}$

Predictive uncertainty: reducible error + irreducible error

Rumsfeld's "Known Unknowns" versus "Unknown Unknowns"

Model Uncertainty

- Entertain a collection of models $\mathcal{M} = \{\mathcal{M}_m,\ m \in M\}$

Model Uncertainty

- Entertain a collection of models $\mathcal{M} = \{\mathcal{M}_m,\ m \in M\}$
- Each model corresponds to a parametric (although possibly infinite-dimensional) distribution of the data **Y**:

$$p_m(\mathbf{y} \mid \boldsymbol{\theta}_m, \mathbf{x}) = p(\mathbf{y} \mid \boldsymbol{\theta}_m, \mathcal{M}_m, \mathbf{x})$$

  where $\boldsymbol{\theta}_m$ corresponds to unknown parameters in the distribution for **Y** under $\mathcal{M}_m$

Model Uncertainty

- Entertain a collection of models $\mathcal{M} = \{\mathcal{M}_m,\ m \in M\}$
- Each model corresponds to a parametric (although possibly infinite-dimensional) distribution of the data $\mathbf{Y}$:

$$p_m(\mathbf{y} \mid \boldsymbol{\theta}_m, \mathbf{x}) = p(\mathbf{y} \mid \boldsymbol{\theta}_m, \mathcal{M}_m, \mathbf{x})$$

  where $\boldsymbol{\theta}_m$ corresponds to unknown parameters in the distribution for $\mathbf{Y}$ under $\mathcal{M}_m$
- Objective: Obtain predictive distributions or summaries at inputs $\mathbf{x}^*$

$$p(\mathbf{y}^* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}^*)$$

Model Uncertainty

- Entertain a collection of models $\mathcal{M} = \{\mathcal{M}_m,\ m \in M\}$
- Each model corresponds to a parametric (although possibly infinite-dimensional) distribution of the data $\mathbf{Y}$:

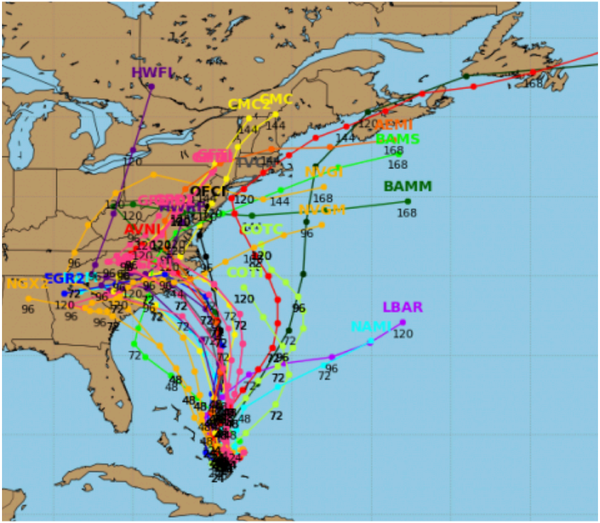
$$p_m(\mathbf{y} \mid \boldsymbol{\theta}_m, \mathbf{x}) = p(\mathbf{y} \mid \boldsymbol{\theta}_m, \mathcal{M}_m, \mathbf{x})$$

  where $\boldsymbol{\theta}_m$ corresponds to unknown parameters in the distribution for $\mathbf{Y}$ under $\mathcal{M}_m$
- Objective: Obtain predictive distributions or summaries at inputs $\mathbf{x}^*$

$$p(\mathbf{y}^* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}^*)$$

  WLOG drop dependence on inputs, $p(\mathbf{y}^* \mid \mathbf{y})$

Multiple Models



Bayesian Perspectives on Model Uncertainty

$\mathcal{M}$-Closed the true data generating model $\mathcal{M}_T$ is one of $\mathcal{M}_m \in M$ but is unknown to researchers

Bayesian Perspectives on Model Uncertainty

$\mathcal{M}$-Closed: the true data generating model $\mathcal{M}_T$ is one of $\mathcal{M}_m \in M$ but is unknown to researchers

$\mathcal{M}$-Complete: the true model $\mathcal{M}_T$ exists but is not included in the model list M. We still wish to use the models in M because of tractability of computations or communication of results, compared with the actual belief model

Bayesian Perspectives on Model Uncertainty

$\mathcal{M}$-Closed: the true data generating model $\mathcal{M}_T$ is one of $\mathcal{M}_m \in M$ but is unknown to researchers

$\mathcal{M}$-Complete: the true model $\mathcal{M}_T$ exists but is not included in the model list M. We still wish to use the models in M because of tractability of computations or communication of results, compared with the actual belief model

$\mathcal{M}$-Open: we know the true model $\mathcal{M}_T$ is not in M, but we cannot specify the explicit form $p(y^* \mid \mathbf{y})$ because it is too difficult conceptually or computationally, we lack time to do so, or do not have the expertise, etc.

Bernardo & Smith (1994), Clyde & Iversen (2013)

Predictive Distributions under $\mathcal{M}$-Closed

- A Bayesian would assign a prior probability, $p(\mathcal{M}_m)$, representing their belief that each model $\mathcal{M}_m$ is the true model.

Predictive Distributions under $\mathcal{M}$-Closed

- A Bayesian would assign a prior probability, $p(\mathcal{M}_m)$, representing their belief that each model $\mathcal{M}_m$ is the true model.
- Distributions $p(\boldsymbol{\theta}_m \mid \mathcal{M}_m)$ characterizing *a priori* uncertainty

Predictive Distributions under $\mathcal{M}$-Closed

- A Bayesian would assign a prior probability, $p(\mathcal{M}_m)$, representing their belief that each model $\mathcal{M}_m$ is the true model.
- Distributions $p(\boldsymbol{\theta}_m \mid \mathcal{M}_m)$ characterizing *a priori* uncertainty
- Bayes Theorem: posterior probability of each model $p(\mathcal{M}_m \mid \mathbf{Y})$

$$
\begin{aligned}
p(\mathcal{M}_m \mid \mathbf{Y}) &= \frac{p(\mathbf{Y}|\mathcal{M}_m)p(\mathcal{M}_m)}{\sum_{m \in M} p(\mathbf{y} \mid \mathcal{M}_m)p(\mathcal{M}_m)}, \quad m \in M \\
\text{where } p(\mathbf{Y} \mid \mathcal{M}_m) &= \int p(\mathbf{Y} \mid \boldsymbol{\theta}_m, \mathcal{M}_m)p(\boldsymbol{\theta}_m \mid \mathcal{M}_m)d\boldsymbol{\theta}_m
\end{aligned}
$$

Predictive Distributions under $\mathcal{M}$-Closed

- A Bayesian would assign a prior probability, $p(\mathcal{M}_m)$, representing their belief that each model $\mathcal{M}_m$ is the true model.
- Distributions $p(\boldsymbol{\theta}_m \mid \mathcal{M}_m)$ characterizing *a priori* uncertainty
- Bayes Theorem: posterior probability of each model $p(\mathcal{M}_m \mid \mathbf{Y})$

$$
\begin{aligned}
p(\mathcal{M}_m \mid \mathbf{Y}) &= \frac{p(\mathbf{Y}|\mathcal{M}_m)p(\mathcal{M}_m)}{\sum_{m \in M} p(\mathbf{y} \mid \mathcal{M}_m)p(\mathcal{M}_m)}, \quad m \in M \\
\text{where } p(\mathbf{Y} \mid \mathcal{M}_m) &= \int p(\mathbf{Y} \mid \boldsymbol{\theta}_m, \mathcal{M}_m)p(\boldsymbol{\theta}_m \mid \mathcal{M}_m)d\boldsymbol{\theta}_m
\end{aligned}
$$

- Predictive distribution

$$
\begin{aligned}
p(\mathbf{y}^*|\mathbf{y}) &= \sum_{m \in \mathcal{M}} p(\mathbf{y}^*|\mathcal{M}_m, \mathbf{y})p(\mathcal{M}_m|\mathbf{y}) \\
&= \sum_{m \in \mathcal{M}} \left[\int p(\mathbf{y}^*|\mathcal{M}_m, \boldsymbol{\theta}_m, \mathbf{y})p(\boldsymbol{\theta}_m|\mathbf{y}, \mathcal{M}_m)\, d\boldsymbol{\theta}_m\right] p(\mathcal{M}_m|\mathbf{y})
\end{aligned}
$$

Estimation and Prediction

Consider the decision problem of estimation/prediction under squared error loss

$$u(Y^*, a) = -(Y^* - a)^2$$

where a is a possible action (u is utility or negative loss) and Y^* is an unknown.

Estimation and Prediction

Consider the decision problem of estimation/prediction under squared error loss

$$u(Y^*, a) = -(Y^* - a)^2$$

where a is a possible action (u is utility or negative loss) and Y^* is an unknown.

From a Bayesian perspective, the solution is to find the action that maximizes expected utility given the observed data $\mathbf{Y}$:

$$\mathsf{E}_{\mathbf{Y}^*|\mathbf{Y}}[u(\mathbf{Y}^*, a)] = -\int (y^* - a)^2 p(\mathbf{y}^* \mid \mathbf{y}) d\mathbf{y}^*$$

where the expectation is taken with respect to the predictive distribution of $\mathbf{Y}^*$ given the observed data $\mathbf{y}$.

Bayesian Model Averaging

- Under the $\mathcal{M}$-closed perspective, optimal solution for prediction is Bayesian Model Averaging

$$a^* = \mathsf{E}_{\mathbf{Y}^*}[\mathbf{Y}^* \mid \mathbf{Y}] = \sum_{m \in \mathcal{M}} p(\mathcal{M}_m \mid \mathbf{Y}) \hat{Y}^*_{\mathcal{M}_m}$$

where $\hat{Y}^*_{\mathcal{M}_m}$ is the predictive mean of $\mathbf{Y}^*$ under model $\mathcal{M}_m$

Bayesian Model Averaging

- Under the $\mathcal{M}$-closed perspective, optimal solution for prediction is Bayesian Model Averaging

$$a^* = \mathsf{E}_{\mathbf{Y}^*}[\mathbf{Y}^* \mid \mathbf{Y}] = \sum_{m \in \mathcal{M}} p(\mathcal{M}_m \mid \mathbf{Y}) \hat{Y}^*_{\mathcal{M}_m}$$

  where $\hat{Y}^*_{\mathcal{M}_m}$ is the predictive mean of $\mathbf{Y}^*$ under model $\mathcal{M}_m$
- Use joint posterior distribution on $\boldsymbol{\theta} \mid \mathcal{M}$ and $\mathcal{M}$ to obtain prediction intervals

Bayesian Model Averaging

- Under the $\mathcal{M}$-closed perspective, optimal solution for prediction is Bayesian Model Averaging

$$a^* = \mathsf{E}_{\mathbf{Y}^*}[\mathbf{Y}^* \mid \mathbf{Y}] = \sum_{m \in \mathcal{M}} p(\mathcal{M}_m \mid \mathbf{Y}) \hat{Y}^*_{\mathcal{M}_m}$$

  where $\hat{Y}^*_{\mathcal{M}_m}$ is the predictive mean of $\mathbf{Y}^*$ under model $\mathcal{M}_m$
- Use joint posterior distribution on $\boldsymbol{\theta} \mid \mathcal{M}$ and $\mathcal{M}$ to obtain prediction intervals
- Full propagation of all “known” uncertainties

Bayesian Model Averaging

- Under the $\mathcal{M}$-closed perspective, optimal solution for prediction is Bayesian Model Averaging

$$a^* = \mathsf{E}_{\mathbf{Y}^*}[\mathbf{Y}^* \mid \mathbf{Y}] = \sum_{m \in \mathcal{M}} p(\mathcal{M}_m \mid \mathbf{Y}) \hat{Y}^*_{\mathcal{M}_m}$$

  where $\hat{Y}^*_{\mathcal{M}_m}$ is the predictive mean of $\mathbf{Y}^*$ under model $\mathcal{M}_m$
- Use joint posterior distribution on $\boldsymbol{\theta} \mid \mathcal{M}$ and $\mathcal{M}$ to obtain prediction intervals
- Full propagation of all "known" uncertainties
- Extensive literature for regression and generalized linear models [Hoeting et al 1999, Clyde & George 2004, Bayarri et al 2012] with invariant priors/Spike & Slab + software
- more complex models via RJ-MCMC, SMC, ABC

Potential Problem with BMA

Two models in $\mathcal{M}$

- $\mathcal{M}_1 : \mathbf{Y} = \mathbf{X}_1 \beta_1 + \mathbf{e}$
- $\mathcal{M}_2 : \mathbf{Y} = \mathbf{X}_2 \beta_2 + \mathbf{e}$

Potential Problem with BMA

Two models in $\mathcal{M}$

- $\mathcal{M}_1 : \mathbf{Y} = \mathbf{X}_1 \beta_1 + \mathbf{e}$
- $\mathcal{M}_2 : \mathbf{Y} = \mathbf{X}_2 \beta_2 + \mathbf{e}$

$$\begin{aligned} \text{True Model} \quad \mathbf{Y} &= \mathbf{X}_1 \beta_{1T} + \mathbf{X}_2 \beta_{2T} + \mathbf{e} \\ \text{BMA} \quad \hat{\mathbf{Y}}^* &= p(\mathcal{M}_1 \mid \mathbf{Y}) \mathbf{X}_1 \hat{\beta}_1 + p(\mathcal{M}_2 \mid \mathbf{Y}) \mathbf{X}_2 \hat{\beta}_2 \end{aligned}$$

Potential Problem with BMA

Two models in $\mathcal{M}$

- $\mathcal{M}_1 : \mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{e}$
- $\mathcal{M}_2 : \mathbf{Y} = \mathbf{X}_2\beta_2 + \mathbf{e}$

$$\begin{array}{rrcl} \text{True Model} & \mathbf{Y} & = & \mathbf{X}_1\beta_{1T} + \mathbf{X}_2\beta_{2T} + \mathbf{e} \\ \text{BMA} & \hat{\mathbf{Y}}^* & = & p(\mathcal{M}_1 \mid \mathbf{Y})\mathbf{X}_1\hat{\beta}_1 + p(\mathcal{M}_2 \mid \mathbf{Y})\mathbf{X}_2\hat{\beta}_2 \end{array}$$

- BMA model weights converge to 1 for the model that is "closest" to true model in Kullback-Leibler divergence

Potential Problem with BMA

Two models in $\mathcal{M}$

- $\mathcal{M}_1 : \mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{e}$
- $\mathcal{M}_2 : \mathbf{Y} = \mathbf{X}_2\beta_2 + \mathbf{e}$

$$
\begin{array}{rrcl}
\text{True Model} & \mathbf{Y} & = & \mathbf{X}_1\beta_{1T} + \mathbf{X}_2\beta_{2T} + \mathbf{e} \\
\text{BMA} & \hat{\mathbf{Y}}^* & = & p(\mathcal{M}_1 \mid \mathbf{Y})\mathbf{X}_1\hat{\beta}_1 + p(\mathcal{M}_2 \mid \mathbf{Y})\mathbf{X}_2\hat{\beta}_2
\end{array}
$$

- BMA model weights converge to 1 for the model that is "closest" to true model in Kullback-Leibler divergence
- BMA only uses predictions from that model

Potential Problem with BMA

Two models in $\mathcal{M}$

- $\mathcal{M}_1 : \mathbf{Y} = \mathbf{X}_1 \beta_1 + \mathbf{e}$
- $\mathcal{M}_2 : \mathbf{Y} = \mathbf{X}_2 \beta_2 + \mathbf{e}$

$$
\begin{array}{rrcl}
\text{True Model} & \mathbf{Y} & = & \mathbf{X}_1 \beta_{1T} + \mathbf{X}_2 \beta_{2T} + \mathbf{e} \\
\text{BMA} & \hat{\mathbf{Y}}^* & = & p(\mathcal{M}_1 \mid \mathbf{Y}) \mathbf{X}_1 \hat{\beta}_1 + p(\mathcal{M}_2 \mid \mathbf{Y}) \mathbf{X}_2 \hat{\beta}_2
\end{array}
$$

- BMA model weights converge to 1 for the model that is "closest" to true model in Kullback-Leibler divergence
- BMA only uses predictions from that model
- In the limit BMA is not consistent if $\mathcal{M}_T \notin \mathcal{M}$

Potential Problem with BMA

Two models in $\mathcal{M}$

- $\mathcal{M}_1 : \mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{e}$
- $\mathcal{M}_2 : \mathbf{Y} = \mathbf{X}_2\beta_2 + \mathbf{e}$

$$\begin{array}{rrcl} \text{True Model} & \mathbf{Y} & = & \mathbf{X}_1\beta_{1T} + \mathbf{X}_2\beta_{2T} + \mathbf{e} \\ \text{BMA} & \hat{\mathbf{Y}}^* & = & p(\mathcal{M}_1 \mid \mathbf{Y})\mathbf{X}_1\hat{\beta}_1 + p(\mathcal{M}_2 \mid \mathbf{Y})\mathbf{X}_2\hat{\beta}_2 \end{array}$$

- BMA model weights converge to 1 for the model that is "closest" to true model in Kullback-Leibler divergence
- BMA only uses predictions from that model
- In the limit BMA is not consistent if $\mathcal{M}_T \notin \mathcal{M}$
- Expand the list of models (prior specification on M)

Potential Problem with BMA

Two models in $\mathcal{M}$

- $\mathcal{M}_1 : \mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{e}$
- $\mathcal{M}_2 : \mathbf{Y} = \mathbf{X}_2\beta_2 + \mathbf{e}$

$$\begin{array}{rrcl} \text{True Model} & \mathbf{Y} & = & \mathbf{X}_1\beta_{1T} + \mathbf{X}_2\beta_{2T} + \mathbf{e} \\ \text{BMA} & \hat{\mathbf{Y}}^* & = & p(\mathcal{M}_1 \mid \mathbf{Y})\mathbf{X}_1\hat{\beta}_1 + p(\mathcal{M}_2 \mid \mathbf{Y})\mathbf{X}_2\hat{\beta}_2 \end{array}$$

- BMA model weights converge to 1 for the model that is "closest" to true model in Kullback-Leibler divergence
- BMA only uses predictions from that model
- In the limit BMA is not consistent if $\mathcal{M}_T \notin \mathcal{M}$
- Expand the list of models (prior specification on M)
- Other model ensembles ?

Combining Models as a Decision Problem

- In $\mathcal{M}$-Complete or $\mathcal{M}$-Open viewpoints, if $\mathcal{M}_T$ is not in the list of models M then $p(\mathcal{M}_m) = 0$ for $\mathcal{M}_m \in M$.

Combining Models as a Decision Problem

- In $\mathcal{M}$-Complete or $\mathcal{M}$-Open viewpoints, if $\mathcal{M}_T$ is not in the list of models M then $p(\mathcal{M}_m) = 0$ for $\mathcal{M}_m \in M$.
- George Box: "All models are wrong, but some may be useful"

Combining Models as a Decision Problem

- In $\mathcal{M}$-Complete or $\mathcal{M}$-Open viewpoints, if $\mathcal{M}_T$ is not in the list of models M then $p(\mathcal{M}_m) = 0$ for $\mathcal{M}_m \in M$.
- George Box: "All models are wrong, but some may be useful"

$$a(\mathbf{w}, \mathbf{Y}) = \sum w_m \hat{Y}^*_m$$

Combining Models as a Decision Problem

- In $\mathcal{M}$-Complete or $\mathcal{M}$-Open viewpoints, if $\mathcal{M}_T$ is not in the list of models M then $p(\mathcal{M}_m) = 0$ for $\mathcal{M}_m \in M$.
- George Box: "All models are wrong, but some may be useful"

$$a(\mathbf{w}, \mathbf{Y}) = \sum w_m \hat{Y}^*_m$$

- Treat weights $\{w_m, m \in m\}$ as part of the action space (rather than an unknown) and maximize posterior expected utility,

$$\mathsf{E}_{\mathbf{Y}^*|\mathbf{y},\mathcal{M}_T}[u(\mathbf{Y}^*, a(\mathbf{w}, \mathbf{y}))] = \int u(\mathbf{y}^*, a(\mathbf{w}, \mathbf{y},))p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)$$

Combining Models as a Decision Problem

- In $\mathcal{M}$-Complete or $\mathcal{M}$-Open viewpoints, if $\mathcal{M}_T$ is not in the list of models M then $p(\mathcal{M}_m) = 0$ for $\mathcal{M}_m \in M$.
- George Box: "All models are wrong, but some may be useful"

$$a(\mathbf{w}, \mathbf{Y}) = \sum w_m \hat{Y}^*_m$$

- Treat weights $\{w_m, m \in m\}$ as part of the action space (rather than an unknown) and maximize posterior expected utility,

$$\mathsf{E}_{\mathbf{Y}^* \mid \mathbf{y}, \mathcal{M}_T}[u(\mathbf{Y}^*, a(\mathbf{w}, \mathbf{y}))] = \int u(\mathbf{y}^*, a(\mathbf{w}, \mathbf{y},))p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)$$

- For negative squared error:

$$-\mathsf{E}_{\mathbf{Y}^* \mid \mathbf{y}, \mathcal{M}_T}\|\mathbf{Y}^* - a(\mathbf{w}, \mathbf{y})\|^2 = -\int \|\mathbf{y}^* - \sum_m w_m \hat{\mathbf{Y}}^*_m\|^2 p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)$$

Combining Models as a Decision Problem

- In $\mathcal{M}$-Complete or $\mathcal{M}$-Open viewpoints, if $\mathcal{M}_T$ is not in the list of models M then $p(\mathcal{M}_m) = 0$ for $\mathcal{M}_m \in M$.
- George Box: "All models are wrong, but some may be useful"

$$a(\mathbf{w}, \mathbf{Y}) = \sum w_m \hat{Y}^*_m$$

- Treat weights $\{w_m, m \in m\}$ as part of the action space (rather than an unknown) and maximize posterior expected utility,

$$\mathsf{E}_{\mathbf{Y}^*|\mathbf{y},\mathcal{M}_T}[u(\mathbf{Y}^*, a(\mathbf{w}, \mathbf{y}))] = \int u(\mathbf{y}^*, a(\mathbf{w}, \mathbf{y},))p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)$$

- For negative squared error:

$$-\mathsf{E}_{\mathbf{Y}^*|\mathbf{y},\mathcal{M}_T}\|\mathbf{Y}^* - a(\mathbf{w}, \mathbf{y})\|^2 = -\int \|\mathbf{y}^* - \sum_m w_m \hat{\mathbf{Y}}^*_m\|^2 p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)$$

- Focus on $\mathcal{M}$-Open case

$\mathcal{M}$-Open Predictive Distribution

- No explicit form for $p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)$ under $\mathcal{M}_T$

$\mathcal{M}$-Open Predictive Distribution

- No explicit form for $p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)$ under $\mathcal{M}_T$
- partition the data: $\mathbf{Y} = (\mathbf{Y}_j, \mathbf{Y}_{(-j)})$

$\mathcal{M}$-Open Predictive Distribution

- No explicit form for $p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)$ under $\mathcal{M}_T$
- partition the data: $\mathbf{Y} = (\mathbf{Y}_j, \mathbf{Y}_{(-j)})$
  - $\mathbf{Y}_j$ is a proxy for $\mathbf{Y}^*$ (the future observation(s))

$\mathcal{M}$-Open Predictive Distribution

- No explicit form for $p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)$ under $\mathcal{M}_T$
- partition the data: $\mathbf{Y} = (\mathbf{Y}_j, \mathbf{Y}_{(-j)})$
  - $\mathbf{Y}_j$ is a proxy for $\mathbf{Y}^*$ (the future observation(s))
  - $\mathbf{Y}_{(-j)}$ is a proxy for $\mathbf{Y}$ (the observed data)

$\mathcal{M}$-Open Predictive Distribution

- No explicit form for $p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)$ under $\mathcal{M}_T$
- partition the data: $\mathbf{Y} = (\mathbf{Y}_j, \mathbf{Y}_{(-j)})$
  - $\mathbf{Y}_j$ is a proxy for $\mathbf{Y}^*$ (the future observation(s))
  - $\mathbf{Y}_{(-j)}$ is a proxy for $\mathbf{Y}$ (the observed data)
- randomly select J partitions,

$$\int u(\mathbf{Y}^*, a(\mathbf{w}, \mathbf{y})) p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)\; d\mathbf{y}^* \approx \frac{1}{J} \sum_{j=1}^{J} u(Y_j, a(\mathbf{w}, Y_{(-j)}))$$

$\mathcal{M}$-Open Predictive Distribution

- No explicit form for $p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)$ under $\mathcal{M}_T$
- partition the data: $\mathbf{Y} = (\mathbf{Y}_j, \mathbf{Y}_{(-j)})$
  - $\mathbf{Y}_j$ is a proxy for $\mathbf{Y}^*$ (the future observation(s))
  - $\mathbf{Y}_{(-j)}$ is a proxy for $\mathbf{Y}$ (the observed data)
- randomly select J partitions,

$$\int u(\mathbf{Y}^*, a(\mathbf{w}, \mathbf{y}))p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)\; d\mathbf{y}^* \approx \frac{1}{J}\sum_{j=1}^{J} u(Y_j, a(\mathbf{w}, Y_{(-j)}))$$

Key et al. + Clyde & Iversen justification of cross-validation to approximate expected posterior utility.

$\mathcal{M}$-Open Predictive Distribution

- No explicit form for $p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)$ under $\mathcal{M}_T$
- partition the data: $\mathbf{Y} = (\mathbf{Y}_j, \mathbf{Y}_{(-j)})$
  - $\mathbf{Y}_j$ is a proxy for $\mathbf{Y}^*$ (the future observation(s))
  - $\mathbf{Y}_{(-j)}$ is a proxy for $\mathbf{Y}$ (the observed data)
- randomly select J partitions,

$$\int u(\mathbf{Y}^*, a(\mathbf{w}, \mathbf{y}))p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_T)\, d\mathbf{y}^* \approx \frac{1}{J}\sum_{j=1}^{J} u(Y_j, a(\mathbf{w}, Y_{(-j)}))$$

  Key et al. + Clyde & Iversen justification of cross-validation to approximate expected posterior utility.
- Guterriez-Pena & Walker approximation to a (limiting) Dirichlet process model for estimating unknown distribution F for $\mathcal{M}_T$

$$\int u(y^*, a^*(\mathbf{w}, \mathbf{y}))dF_n(y^*) \to \frac{1}{n}\sum_{i=1}^{n} u(y_i, a^*(\mathbf{w}, \mathbf{Y}_{(-i)}))$$

Optimization Problem under Approximation

Find weights

$$\hat{w} = \arg\max_{\mathbf{w}} \; -\frac{1}{J}\sum_{j=1}^{J}\left(Y_j - \sum_{m\in\mathcal{M}} w_m \hat{Y}_{(-j),\mathcal{M}_m}\right)^2$$

Optimization Problem under Approximation

Find weights

$$\hat{w} = \arg\max_{\mathbf{w}} \; -\frac{1}{J}\sum_{j=1}^{J}\left(Y_j - \sum_{m\in\mathcal{M}} w_m \hat{Y}_{(-j),\mathcal{M}_m}\right)^2$$

Constrained Solution:

$$\begin{aligned}
\text{Find weights: } \hat{w} \quad &= \quad \arg\max_{\mathbf{w}} \; -\frac{1}{J}\sum_{j=1}^{J}\left(Y_j - \sum_{m\in\mathcal{M}} w_m \hat{Y}_{(-j),\mathcal{M}_m}\right)^2 \\
\text{subject to } \sum_i w_m \quad &= \quad 1 \\
w_m \quad &\geq \quad 0 \quad \forall m \in \mathcal{M}
\end{aligned}$$

Optimization Problem under Approximation

Find weights

$$\hat{w} = \arg\max_{\mathbf{w}} \; -\frac{1}{J}\sum_{j=1}^{J}\left(Y_j - \sum_{m\in\mathcal{M}} w_m \hat{Y}_{(-j),\mathcal{M}_m}\right)^2$$

Constrained Solution:

$$\begin{aligned}
\text{Find weights: } \hat{w} &= \arg\max_{\mathbf{w}} \; -\frac{1}{J}\sum_{j=1}^{J}\left(Y_j - \sum_{m\in\mathcal{M}} w_m \hat{Y}_{(-j),\mathcal{M}_m}\right)^2 \\
\text{subject to } \sum_{i} w_m &= 1 \\
w_m &\geq 0 \quad \forall m \in \mathcal{M}
\end{aligned}$$

Equivalent representation (Lagrangian):

$$-\frac{1}{J}\sum_{j=1}^{J}\left(Y_j - \sum_{m\in\mathcal{M}} w_m \hat{Y}_{(-j),\mathcal{M}_m}\right)^2 - \lambda_0\left(\sum_{m}^{M} w_m - 1\right) + \sum_{m}^{M} \lambda_m w_m$$

Comments on Solutions

Let $\mathbf{e} = [e]_{ji} = Y_j - \hat{Y}_{(-j)\mathcal{M}_i}$ denote the $n \times M$ matrix of residuals for predicting Y_j under model $\mathcal{M}_i$.

Comments on Solutions

Let $\mathbf{e} = [e]_{ji} = Y_j - \hat{Y}_{(-j)\mathcal{M}_i}$ denote the $n \times M$ matrix of residuals for predicting Y_j under model $\mathcal{M}_i$.

- With sum to 1 constraint alone, $\hat{\mathbf{w}} \propto (\mathbf{e}^T\mathbf{e})^{-1}\mathbf{1}$

Comments on Solutions

Let $\mathbf{e} = [e]_{ji} = Y_j - \hat{Y}_{(-j)\mathcal{M}_i}$ denote the $n \times M$ matrix of residuals for predicting Y_j under model $\mathcal{M}_i$.

- With sum to 1 constraint alone, $\hat{\mathbf{w}} \propto (\mathbf{e}^T\mathbf{e})^{-1}\mathbf{1}$
- If residuals from models are uncorrelated, then weights are proportional to the inverse of the cross-validation MSE for model $\mathcal{M}_i$, $\text{MSE}_i = \sum_j e_{ij}^2$

Comments on Solutions

Let $\mathbf{e} = [e]_{ji} = Y_j - \hat{Y}_{(-j)\mathcal{M}_i}$ denote the $n \times M$ matrix of residuals for predicting Y_j under model $\mathcal{M}_i$.

- With sum to 1 constraint alone, $\hat{\mathbf{w}} \propto (\mathbf{e}^T\mathbf{e})^{-1}\mathbf{1}$
- If residuals from models are uncorrelated, then weights are proportional to the inverse of the cross-validation MSE for model $\mathcal{M}_i$, $\text{MSE}_i = \sum_j e_{ij}^2$
- With highly correlated predictions/residual weights may be negative and highly unstable

Comments on Solutions

Let $\mathbf{e} = [e]_{ji} = Y_j - \hat{Y}_{(-j)\mathcal{M}_i}$ denote the $n \times M$ matrix of residuals for predicting Y_j under model $\mathcal{M}_i$.

- With sum to 1 constraint alone, $\hat{\mathbf{w}} \propto (\mathbf{e}^T\mathbf{e})^{-1}\mathbf{1}$
- If residuals from models are uncorrelated, then weights are proportional to the inverse of the cross-validation MSE for model $\mathcal{M}_i$, $\text{MSE}_i = \sum_j e_{ij}^2$
- With highly correlated predictions/residual weights may be negative and highly unstable
- Non-negativity lasso-like constraint stabilizes weights, and may drive weights to 0 for similar models

Comments on Solutions

Let $\mathbf{e} = [e]_{ji} = Y_j - \hat{Y}_{(-j)\mathcal{M}_i}$ denote the $n \times M$ matrix of residuals for predicting Y_j under model $\mathcal{M}_i$.

- With sum to 1 constraint alone, $\hat{\mathbf{w}} \propto (\mathbf{e}^T\mathbf{e})^{-1}\mathbf{1}$
- If residuals from models are uncorrelated, then weights are proportional to the inverse of the cross-validation MSE for model $\mathcal{M}_i$, $\text{MSE}_i = \sum_j e_{ij}^2$
- With highly correlated predictions/residual weights may be negative and highly unstable
- Non-negativity lasso-like constraint stabilizes weights, and may drive weights to 0 for similar models

Provides a Bayesian justification for classical stacking (Wolpert 1992, Breiman 1996)

Comments on Solutions

Let $\mathbf{e} = [e]_{ji} = Y_j - \hat{Y}_{(-j)\mathcal{M}_i}$ denote the $n \times M$ matrix of residuals for predicting Y_j under model $\mathcal{M}_i$.

- With sum to 1 constraint alone, $\hat{\mathbf{w}} \propto (\mathbf{e}^T\mathbf{e})^{-1}\mathbf{1}$
- If residuals from models are uncorrelated, then weights are proportional to the inverse of the cross-validation MSE for model $\mathcal{M}_i$, $\text{MSE}_i = \sum_j e_{ij}^2$
- With highly correlated predictions/residual weights may be negative and highly unstable
- Non-negativity lasso-like constraint stabilizes weights, and may drive weights to 0 for similar models

Provides a Bayesian justification for classical stacking (Wolpert 1992, Breiman 1996)

Super-Learners! `h2oEnsemble`

Ovarian Cancer Example

Predict short *vs.* long-term survival given primary tumor's molecular phenotype.

Ovarian Cancer Example

Predict short *vs.* long-term survival given primary tumor's molecular phenotype.

- Retrospective sample of survivors of advanced stage serous ovarian cancer
  - $n = 30$ short-term (< 3 years)
  - $n = 24$ long-term (> 7 years)

Ovarian Cancer Example

Predict short *vs.* long-term survival given primary tumor's molecular phenotype.

- Retrospective sample of survivors of advanced stage serous ovarian cancer
  - $n = 30$ short-term (< 3 years)
  - $n = 24$ long-term (> 7 years)
- Eleven early stage (I/II) cases for external validation.

Ovarian Cancer Example

Predict short *vs.* long-term survival given primary tumor's molecular phenotype.

- Retrospective sample of survivors of advanced stage serous ovarian cancer
  - $n = 30$ short-term $(< 3$ years$)$
  - $n = 24$ long-term $(> 7$ years$)$
- Eleven early stage (I/II) cases for external validation.
- Affymetrix U133a expression microarray; $22,283$ genes.

Ovarian Cancer Example

Predict short *vs.* long-term survival given primary tumor's molecular phenotype.

- Retrospective sample of survivors of advanced stage serous ovarian cancer
  - $n = 30$ short-term $(< 3$ years$)$
  - $n = 24$ long-term $(> 7$ years$)$
- Eleven early stage (I/II) cases for external validation.
- Affymetrix U133a expression microarray; $22,283$ genes.
- 6 clinical variables

Ovarian Cancer Example

Predict short *vs.* long-term survival given primary tumor's molecular phenotype.

- Retrospective sample of survivors of advanced stage serous ovarian cancer
  - $n = 30$ short-term (< 3 years)
  - $n = 24$ long-term (> 7 years)
- Eleven early stage (I/II) cases for external validation.
- Affymetrix U133a expression microarray; $22,283$ genes.
- 6 clinical variables

Models for M-Open Model Averaging (MOMA)

Three classes of models:

- Clinical Trees: (5 models) Prospective classification and regression tree models using only clinical variables such as age, post-treatment CA125 levels, etc.

Models for M-Open Model Averaging (MOMA)

Three classes of models:

- Clinical Trees: (5 models) Prospective classification and regression tree models using only clinical variables such as age, post-treatment CA125 levels, etc.
- Expression Trees: (4 models) Prospective Classification and regression tree models using only expression data.

Models for M-Open Model Averaging (MOMA)

Three classes of models:

- Clinical Trees: (5 models) Prospective classification and regression tree models using only clinical variables such as age, post-treatment CA125 levels, etc.
- Expression Trees: (4 models) Prospective Classification and regression tree models using only expression data.
- Expression LDA: (4 models) Retrospective discriminant models built using expression data given survival.

Models for M-Open Model Averaging (MOMA)

Three classes of models:

- Clinical Trees: (5 models) Prospective classification and regression tree models using only clinical variables such as age, post-treatment CA125 levels, etc.
- Expression Trees: (4 models) Prospective Classification and regression tree models using only expression data.
- Expression LDA: (4 models) Retrospective discriminant models built using expression data given survival.

Find MOMA ($\mathcal{M}$-Open Model Averaging) weights $\hat{w}$ using all long term and short term survivors

Models for M-Open Model Averaging (MOMA)

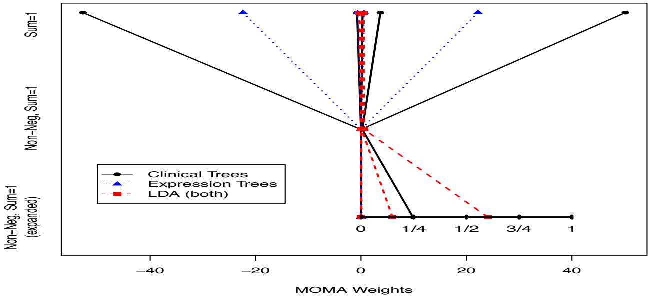
Three classes of models:

- Clinical Trees: (5 models) Prospective classification and regression tree models using only clinical variables such as age, post-treatment CA125 levels, etc.
- Expression Trees: (4 models) Prospective Classification and regression tree models using only expression data.
- Expression LDA: (4 models) Retrospective discriminant models built using expression data given survival.

Find MOMA ($\mathcal{M}$-Open Model Averaging) weights $\hat{w}$ using all long term and short term survivors

- sum to 1 constraint
- $+$ non-negativity constraint

MOMA Weights



Validation Experiment

- 5-fold cross validation; 5 splits of data into two groups: Training $\mathbf{Y}$ and Validation $\mathbf{Y}^*$

Validation Experiment

- 5-fold cross validation; 5 splits of data into two groups: Training $\mathbf{Y}$ and Validation $\mathbf{Y}^*$
- Use training data to obtain model weights $\hat{\mathbf{w}}$ via LOO

Validation Experiment

- 5-fold cross validation; 5 splits of data into two groups: Training $\mathbf{Y}$ and Validation $\mathbf{Y}^*$
- Use training data to obtain model weights $\hat{\mathbf{w}}$ via LOO
- Construct $\mathcal{M}$-Open Model Averaging (MOMA) estimates of probability of long term survival $\hat{p}_j = \sum_i \hat{w}_i \hat{Y}^*_{\mathcal{M}_i}(\mathbf{Y})$ for validation samples

Validation Experiment

- 5-fold cross validation; 5 splits of data into two groups: Training $\mathbf{Y}$ and Validation $\mathbf{Y}^*$
- Use training data to obtain model weights $\hat{\mathbf{w}}$ via LOO
- Construct $\mathcal{M}$-Open Model Averaging (MOMA) estimates of probability of long term survival $\hat{p}_j = \sum_i \hat{w}_i \hat{Y}^*_{\mathcal{M}_i}(\mathbf{Y})$ for validation samples
- Classify as Long Term Survivor $\hat{p}_j \geq 1/2$

Validation Experiment

- 5-fold cross validation; 5 splits of data into two groups: Training $\mathbf{Y}$ and Validation $\mathbf{Y}^*$
- Use training data to obtain model weights $\hat{\mathbf{w}}$ via LOO
- Construct $\mathcal{M}$-Open Model Averaging (MOMA) estimates of probability of long term survival $\hat{p}_j = \sum_i \hat{w}_i \hat{Y}^*_{\mathcal{M}_i}(\mathbf{Y})$ for validation samples
- Classify as Long Term Survivor $\hat{p}_j \geq 1/2$
- Compute classification accuracy over 5 Splits

MOMA with Sum-to-1 Constraint

	set1	set2	set3	set4	set5
clin1	53.08	−4.43	−0.01	−24.41	15.94
clin2	−79.92	−5.16	0.90	0.80	−4.63
clin3	−1.25	−0.24	−0.90	−0.01	5.35
clin4	27.36	10.14	−0.33	23.73	−17.24
clin5	1.13	0.27	0.27	0.36	0.55
tree1	−0.05	−0.55	−2.92	0.03	27.93
tree2	−0.12	−0.07	−3.21	−0.62	0.63
tree3	0.51	0.53	0.15	0.48	−3.35
tree4	−0.28	0.22	6.26	−0.04	−24.10
lda100.P1	−0.40	0.04	−0.01	0.02	−0.11
lda100.P2	0.44	−0.02	0.53	−0.06	−0.07
lda200.P1	0.30	0.17	−0.32	0.09	−0.03
lda200.P2	0.21	0.08	0.60	0.63	0.12
Accuracy	0.64	0.64	0.46	0.73	0.60

MOMA with Non-negativity Constraint

	set1	set2	set3	set4	set5
clin1	0.00	0.07	0.00	0.00	0.00
clin2	0.00	0.00	0.00	0.00	0.00
clin3	0.00	0.00	0.00	0.00	0.00
clin4	0.00	0.11	0.00	0.00	0.00
clin5	0.30	0.17	0.07	0.41	0.00
tree1	0.00	0.00	0.00	0.00	0.77
tree2	0.00	0.00	0.00	0.00	0.21
tree3	0.23	0.44	0.21	0.00	0.01
tree4	0.00	0.00	0.00	0.00	0.01
lda100.P1	0.00	0.00	0.00	0.00	0.00
lda100.P2	0.22	0.00	0.30	0.00	0.00
lda200.P1	0.00	0.00	0.00	0.00	0.00
lda200.P2	0.26	0.21	0.41	0.58	0.00
Accuracy	0.82	0.73	0.55	0.73	0.60

More General Utilities - Yao et al (2018)

- Under negative squared error loss, only have optimal point predictions

More General Utilities - Yao et al (2018)

- Under negative squared error loss, only have optimal point predictions
- Stacking probabilistic forecasts $P \in \mathcal{P}$

$$a(\mathbf{w}, \mathbf{y}) = \sum_m w_m p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_m)$$

More General Utilities - Yao et al (2018)

- Under negative squared error loss, only have optimal point predictions
- Stacking probabilistic forecasts $P \in \mathcal{P}$

$$a(\mathbf{w}, \mathbf{y}) = \sum_m w_m p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_m)$$

- Proper Scoring rules: $S(Q, Q) \geq S(P, Q)$ for $P, Q \in \mathcal{P}$

$$S(P, Q) \equiv \int S(P, \omega) dQ(\omega)$$

More General Utilities - Yao et al (2018)

- Under negative squared error loss, only have optimal point predictions
- Stacking probabilistic forecasts $P \in \mathcal{P}$

$$a(\mathbf{w}, \mathbf{y}) = \sum_m w_m p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_m)$$

- Proper Scoring rules: $S(Q, Q) \geq S(P, Q)$ for $P, Q \in \mathcal{P}$

$$S(P, Q) \equiv \int S(P, \omega) dQ(\omega)$$

- Strictly Proper $S(Q, Q) \geq S(P, Q)$ with equality only when $P = Q$ almost surely

More General Utilities - Yao et al (2018)

- Under negative squared error loss, only have optimal point predictions
- Stacking probabilistic forecasts $P \in \mathcal{P}$

$$a(\mathbf{w}, \mathbf{y}) = \sum_m w_m p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_m)$$

- Proper Scoring rules: $S(Q, Q) \geq S(P, Q)$ for $P, Q \in \mathcal{P}$

$$S(P, Q) \equiv \int S(P, \omega) dQ(\omega)$$

- Strictly Proper $S(Q, Q) \geq S(P, Q)$ with equality only when $P = Q$ almost surely
- Negative Quadratic Loss is proper, but not a strictly proper scoring rule

More General Utilities - Yao et al (2018)

- Under negative squared error loss, only have optimal point predictions
- Stacking probabilistic forecasts $P \in \mathcal{P}$

$$a(\mathbf{w}, \mathbf{y}) = \sum_m w_m p(\mathbf{y}^* \mid \mathbf{y}, \mathcal{M}_m)$$

- Proper Scoring rules: $S(Q, Q) \geq S(P, Q)$ for $P, Q \in \mathcal{P}$

$$S(P, Q) \equiv \int S(P, \omega) dQ(\omega)$$

- Strictly Proper $S(Q, Q) \geq S(P, Q)$ with equality only when $P = Q$ almost surely
- Negative Quadratic Loss is proper, but not a strictly proper scoring rule
- Logarithmic Score $S(P, y^*) = \log(p(y^*))$

Probabilistic Forecasting

- Ensemble BMA (Raftery + coauthors 2005 . . .) weather forecasting

Probabilistic Forecasting

- Ensemble BMA (Raftery + coauthors 2005 . . .) weather forecasting

$$\arg\max_{\mathbf{w},\sigma^2} \sum_i \log\left(\sum_m^M w_m p(y_i^* \mid \mathbf{y}, \mathcal{M}_m, \sigma^2\right)$$

Probabilistic Forecasting

- Ensemble BMA (Raftery + coauthors 2005 ...) weather forecasting

$$\arg\max_{\mathbf{w},\sigma^2} \sum_i \log\Big(\sum_m^M w_m p(y_i^* \mid \mathbf{y}, \mathcal{M}_m, \sigma^2)\Big)$$

  - p_m Gaussian distributions centered at $a_m + b_m \hat{\mathbf{Y}}_m^*$

Probabilistic Forecasting

- Ensemble BMA (Raftery + coauthors 2005 ...) weather forecasting

$$\arg\max_{\mathbf{w},\sigma^2} \sum_i \log\left(\sum_m^M w_m p(y_i^* \mid \mathbf{y}, \mathcal{M}_m, \sigma^2\right)$$

  - p_m Gaussian distributions centered at $a_m + b_m \hat{\mathbf{Y}}_m^*$
  - allows for bias and calibration of computer model output

Probabilistic Forecasting

- Ensemble BMA (Raftery + coauthors 2005 . . .) weather forecasting

$$\arg\max_{\mathbf{w},\sigma^2} \sum_i \log(\sum_m^M w_m p(y_i^* \mid \mathbf{y}, \mathcal{M}_m, \sigma^2)$$

  - p_m Gaussian distributions centered at $a_m + b_m \hat{\mathbf{Y}}_m^*$
  - allows for bias and calibration of computer model output
  - common unknown variance σ^2 in each component

Probabilistic Forecasting

- Ensemble BMA (Raftery + coauthors 2005 ...) weather forecasting

$$\arg\max_{\mathbf{w},\sigma^2} \sum_i \log\Big(\sum_m^M w_m p(y_i^* \mid \mathbf{y}, \mathcal{M}_m, \sigma^2)\Big)$$

  - p_m Gaussian distributions centered at $a_m + b_m \hat{\mathbf{Y}}_m^*$
  - allows for bias and calibration of computer model output
  - common unknown variance σ^2 in each component
  - weights evolve with time

Probabilistic Forecasting

- Ensemble BMA (Raftery + coauthors 2005 ...) weather forecasting

$$\arg\max_{\mathbf{w},\sigma^2} \sum_i \log(\sum_m^M w_m p(y_i^* \mid \mathbf{y}, \mathcal{M}_m, \sigma^2)$$

  - p_m Gaussian distributions centered at $a_m + b_m \hat{\mathbf{Y}}_m^*$
  - allows for bias and calibration of computer model output
  - common unknown variance σ^2 in each component
  - weights evolve with time
  - multivariate outcomes
- West + coauthors Dynamic Linear Models (economic forecasting) with dynamic weights

Probabilistic Forecasting

- Ensemble BMA (Raftery + coauthors 2005 ...) weather forecasting

$$\arg\max_{\mathbf{w},\sigma^2} \sum_i \log\left(\sum_m^M w_m p(y_i^* \mid \mathbf{y}, \mathcal{M}_m, \sigma^2)\right)$$

  - p_m Gaussian distributions centered at $a_m + b_m \hat{\mathbf{Y}}_m^*$
  - allows for bias and calibration of computer model output
  - common unknown variance σ^2 in each component
  - weights evolve with time
  - multivariate outcomes
- West + coauthors Dynamic Linear Models (economic forecasting) with dynamic weights
- Gaussian Process emulators for computer models and statistical models

Discussion

- Model Averaging for Uncertainty Quantification under different perspectives

Discussion

- Model Averaging for Uncertainty Quantification under different perspectives
- Dependence on Choice of Utility Functions
  - squared error loss - point estimates
  - proper scoring rules - distributions
  - quantiles

Discussion

- Model Averaging for Uncertainty Quantification under different perspectives
- Dependence on Choice of Utility Functions
  - squared error loss - point estimates
  - proper scoring rules - distributions
  - quantiles
- Partitions of data for approximation? LOO, k-fold, sequential

Discussion

- Model Averaging for Uncertainty Quantification under different perspectives
- Dependence on Choice of Utility Functions
  - squared error loss - point estimates
  - proper scoring rules - distributions
  - quantiles
- Partitions of data for approximation? LOO, k-fold, sequential
- Incorporation of Model Complexity/Regularization in Utility (sum to one?)

Discussion

- Model Averaging for Uncertainty Quantification under different perspectives
- Dependence on Choice of Utility Functions
  - squared error loss - point estimates
  - proper scoring rules - distributions
  - quantiles
- Partitions of data for approximation? LOO, k-fold, sequential
- Incorporation of Model Complexity/Regularization in Utility (sum to one?)
- Optimization: Quadratic programming, EM, variational, ABC

Discussion

- Model Averaging for Uncertainty Quantification under different perspectives
- Dependence on Choice of Utility Functions
  - squared error loss - point estimates
  - proper scoring rules - distributions
  - quantiles
- Partitions of data for approximation? LOO, k-fold, sequential
- Incorporation of Model Complexity/Regularization in Utility (sum to one?)
- Optimization: Quadratic programming, EM, variational, ABC
- Mixture Models and Mixtures of Experts

Discussion

- Model Averaging for Uncertainty Quantification under different perspectives
- Dependence on Choice of Utility Functions
  - squared error loss - point estimates
  - proper scoring rules - distributions
  - quantiles
- Partitions of data for approximation? LOO, k-fold, sequential
- Incorporation of Model Complexity/Regularization in Utility (sum to one?)
- Optimization: Quadratic programming, EM, variational, ABC
- Mixture Models and Mixtures of Experts
- SAMSI Program 2018-19 Model Uncertainty and Uncertainty Quantification