Model complexity and comparison

Vadim Strijov,
Visiting Professor at IAM METU

Computing Center of the Russian Academy of Sciences

Institute of Applied Mathematics,
Middle East Technical University
June 8th, 2012

The plan

1. Coherent Bayesian inference.
2. Evidence of models.
3. Model comparison.

The most probable parameters

$$\mathbf{w}_{\text{MP}} = \arg\max_{\mathbf{w}\in\mathcal{W}} p(\mathbf{w}|D, f, A, B),$$

of the model f are estimated using the Bayesian approach

$$p(\mathbf{w}|D, f, A, B) = \frac{p(D|\mathbf{w}, f, B)p(\mathbf{w}|f, A)}{\int p(D|\mathbf{w}', f, B)p(\mathbf{w}'|f, A)d\mathbf{w}'}.$$

The likelihood function $p(D|\mathbf{w}, f, B)$ is defined by the hypothesis of distribution of the dependent variable $\mathbf{y}$.
The model evidence

$$\mathcal{E}\left(f(\mathbf{w}, \mathbf{x})\right) = \int p(D|\mathbf{w}, f, B)p(\mathbf{w}|f, A)d\mathbf{w}.$$

There given:

- the sample set D,
- the split of the sample index set $\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$ into the learning and test subsets,
- the finite set of models $\mathcal{F} = \{f_k | k \in \mathcal{K}\}$,
- the error function S (defined by the data generation hypothesis $S = -\ln\bigl(p(D|\mathbf{w}, B, f)\bigr)$, or by some practical considerations).

One must select a model f_{k^*} index k^* such that

$$k^* = \arg\min_{k \in \mathcal{K}} S(f_k|\hat{\mathbf{w}}_k, D_{\mathcal{T}}),$$

where the parameters $\hat{\mathbf{w}}_k$ estimated as either most probable or most likely

$$\hat{\mathbf{w}}_k = \arg\min_{\mathbf{w}_k \in \mathcal{W}} S(\mathbf{w}_k|f_k, D_{\mathcal{L}}).$$

There given:

- the sample set D,
- the finite set of models $\mathcal{F} = \{f_k | k \in \mathcal{K}\}$.

One must select the most evident model f_{k^*}, such that

$$k^* = \arg\max_{k \in \mathcal{K}} p(f_k|D) = \arg\max_{k \in \mathcal{K}} \int_{\mathbf{w} \in \mathcal{W}} p(D|\mathbf{w}, B, f_k) p(\mathbf{w}|D, A, f_k) d\mathbf{w}.$$

If we assume the prior probabilities of models are equal,

$$p(f_1) = p(f_2) = \cdots = p(f_K),$$

then the most evident model selection problem is stated as the most probable model selection problem.

There given:

- the sample set D, the model $f = f(\mathbf{w}, \mathbf{x})$,
- the data generation hypothesis, it defines the error function

$$S(\mathbf{w}) = -\ln\big(p(D|\mathbf{w}, B, f)p(\mathbf{w}|A, f)\big).$$

One must estimate the most probable parameters $\mathbf{w}_{\text{MP}}$

$$\mathbf{w}_{\text{MP}} = \arg\min_{\mathbf{w}\in\mathcal{W}} S(\mathbf{w}, D, \hat{A}, \hat{B}, f).$$

One must estimate corresponding hyperparameters A, B

$$\hat{A}, \hat{B} = \arg\min_{A,B} \Phi\big(S(\mathbf{w}_{\text{MP}}, D, A, B, f)\big).$$

Maximize the model evidence $p(D|A,\beta)$ according to A and β

$$p(D|A,\beta) = \int p(D|\mathbf{w},A,\beta)p(\mathbf{w}|A)d\mathbf{w} \to \max.$$

Use the Laplace approximation,

$$p(D|A,\beta) = \frac{1}{Z_{\mathbf{w}}(A)}\frac{1}{Z_D(\beta)}\int \exp(-S(\mathbf{w}))d\mathbf{w}.$$

Substitute $Z_{\mathbf{w}}(A)$, $Z_D(\beta)$ and $S(\mathbf{w})$ and find the logarithm of it:

$$p(D|A,\beta) = \frac{1}{Z_{\mathbf{w}}(A)}\frac{1}{Z_D(\beta)}\exp(-S(\mathbf{w}_0))(2\pi)^{\frac{n}{2}}|H|^{-\frac{1}{2}}.$$

$$\ln p(D|A,\beta) = \underbrace{-\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|A|}_{Z_{\mathbf{w}}^{-1}(A)}\underbrace{-\frac{m}{2}\ln 2\pi + \frac{m}{2}\ln\beta}_{Z_D^{-1}(\beta)}\underbrace{-S(\mathbf{w}_0) + \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|H|}_{Z_S} =$$

$$= -\frac{1}{2}\ln|A| - \frac{m}{2}\ln 2\pi + \frac{m}{2}\ln\beta\underbrace{-\beta E_D - E_{\mathbf{w}}}_{-S(\mathbf{w}_0)} - \frac{1}{2}\ln|H|.$$

Solve the optimization problems

$$\frac{\partial}{\partial A} \ln p(D|A,\beta) = 0 \quad \text{and}$$

$$\frac{\partial}{\partial \beta} \ln p(D|A,\beta) = 0.$$

As the result of the evidence maximization we obtain

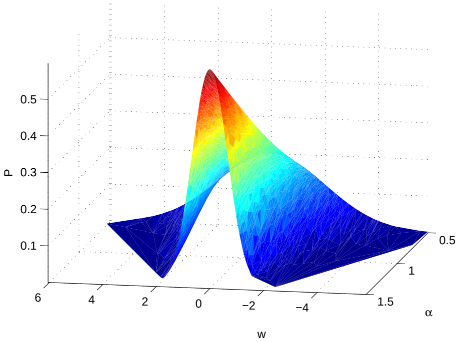
$$2\alpha_j E'_{\mathbf{w}} = n - \gamma_j, \quad \text{where} \quad \gamma_j = \frac{\alpha_j}{\lambda_j + \alpha_j} \quad \text{and}$$

$$2\beta E'_D = m - \sum_{j=1}^{n} \gamma_j.$$

Estimate the hyperparameters α and β_i iteratively,

$$\alpha_j^{\text{new}} = \frac{n - \gamma_j}{2E'_{\mathbf{w}}}, \qquad \beta^{\text{new}} = \frac{m - \sum\limits_{j=1}^{n} \gamma_j}{2E'_D}.$$

How the distribution of parameters depends on $A = \alpha I_n$



- z-axis: $p(\mathbf{w}|D, f, A, B)$ the distribution of parameters,
- y-axis: α the inverted covariance,
- x-axis: w the model parameter.

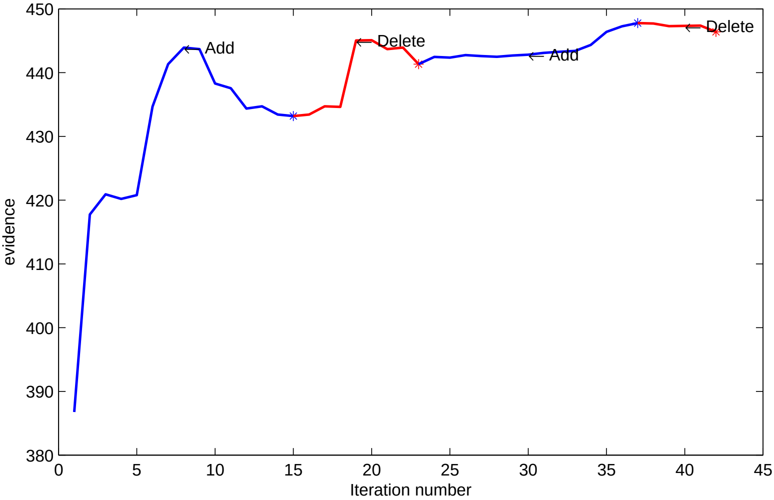
One must find the feature indexes $\mathcal{A} \subseteq \mathcal{J}$.

Step 0. $\mathcal{A}_s = \emptyset$.

Step s.

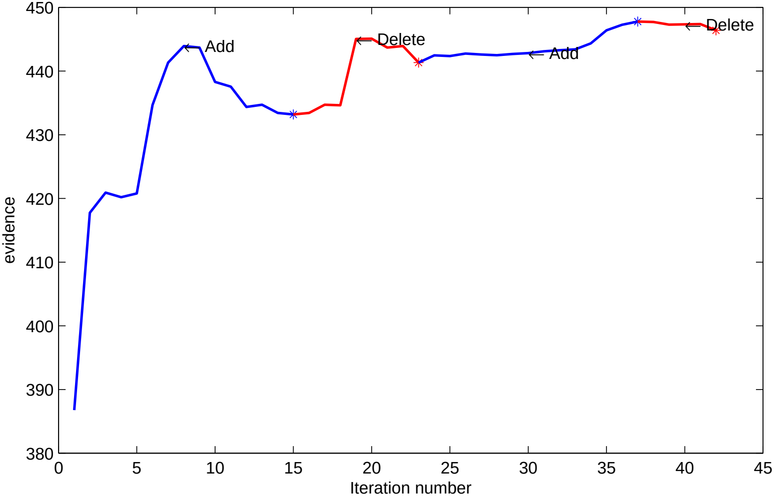
1. **Add** the next feature $\mathcal{A}' = \mathcal{A} \cup \{j\}$, where $j \in \mathcal{J} \setminus \mathcal{A}$, according to a predefined criterion (max correlation or min VIF) until $\mathcal{E}\left(f(\mathbf{w}'_{\mathcal{A}}, \mathbf{x})\right)$ decreases.
2. **Delete** the most informative features $A' = A \setminus \{j\}$, where $j \in \mathcal{A}$, according to the variances $A = \text{diag}(\alpha_1, \dots \alpha_{|\mathcal{A}|})$ until $\mathcal{E}\left(f(\mathbf{w}'_{\mathcal{A}}, \mathbf{x})\right)$ decreases.

- Iterate until convergency of $\mathcal{E}$.

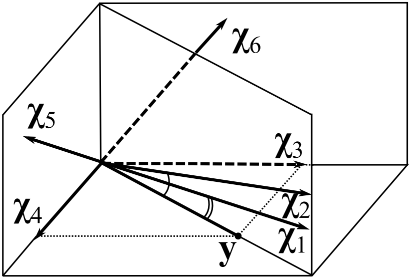


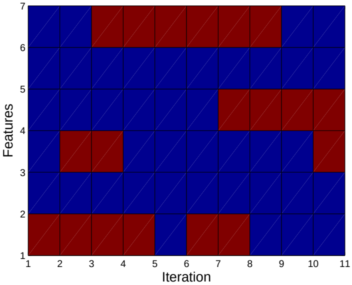
Add and **Delete** features until the evidence goes down.

Model selection by evidence maximization, an example



Add and **Delete** features, until the evidence goes down.





The red color means the feature is included into the active set $\mathcal{A}$.

Ventia non sunt multiplicanda praeter necessitatem.

Occam's razor: entities (model elements) must not be multiplied beyond necessity.

Coherent Bayesian Inference is a method of the model comparison.
This method uses Bayesian inference two times:

1. to estimate the posterior probability of the model parameters and
2. to estimate the posterior probability of the model itself.

Consider a finite set of models $f_1, \dots, f_M$ that fit the data D. Denote prior probability of i-th model by $p(f_i)$. After the data have come, the posterior probability of the model

$$p(f_i|D) = \frac{p(D|f_i)p(f_i)}{\sum_{j=1}^{M} p(D|f_j)p(f_j)}.$$

The probability $p(D|f_i)$ of data D, given model f_i is called the evidence of the model f_i.
Since the denominator for all models from the set is the same,

$$p(D) = \sum_{j=1}^{n} p(D|f_j)p(f_j),$$

then

$$\frac{p(f_i|D)}{p(f_j|D)} = \frac{p(f_i)p(D|f_i)}{p(f_j)p(D|f_j)}.$$

Assume the prior probabilities to be equal, $p(f_i) = p(f_j)$.

Let there be given the series $\{-1, 3, 7, 11\}$. One must forecast the next two elements.
The model f_a:

$$x_{i+1} = x_i + 4$$

gives the next elements $15, 19$.
The model f_c:

$$x_{i+1} = -\frac{x_i^3}{11} + \frac{9x_i^2}{11} + \frac{23}{11}$$

gives the next elements $-19.9, 1043.8$.

Let the prior probabilities be equal or comparable.
Let each parameter of the models is in the set

$$\{-50, \dots, 0, \dots, 50\}.$$

The parameters $(n = 4, x_1 = -1)$ brings the proper model with zero-error.
The evidence of the model f_a is

$$p(D|f_a) = \frac{1}{101}\frac{1}{101} = 0.00010.$$

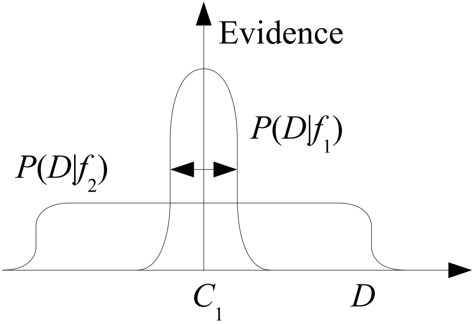
Let the denominators of the second models are in the set $\{0, \dots, 50\}$.
Take account of $c = -1/11 = -2/22 = -3/33 = -4/44$.
The evidence of the model f_c is

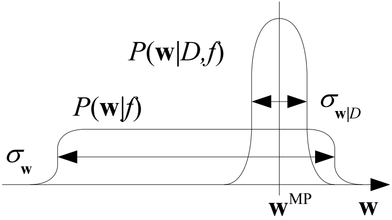
$$p(D|f_c) = \left(\frac{1}{101}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{4}{101}\frac{1}{50}\right) = 4.9202\ldots\times 10^{-12}.$$

The result of the model comparison is

$$\frac{p(D|f_a)}{p(D|f_c)} = \frac{0.00010}{2.5\times 10^{-12}}.$$

If f_2 — is more complex model, then its distribution $p(D|f_2)$ has smaller values (variance has greater values). If the errors of both models are equal, then the simple model f_1 is more probable than the complex model f_2.





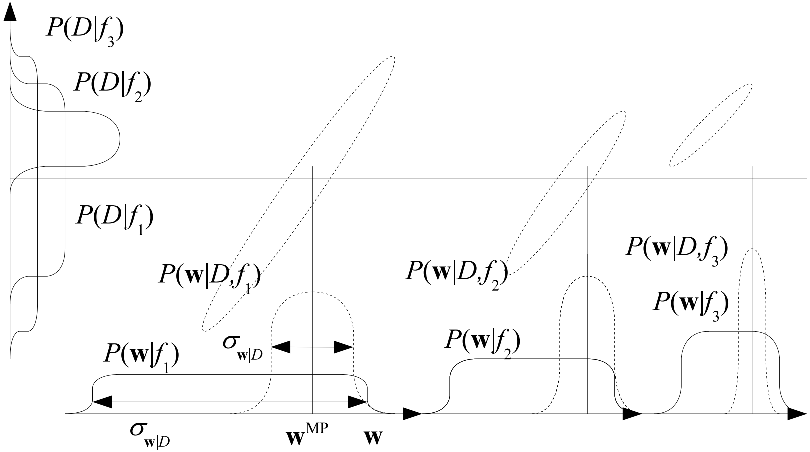
The Occam factor is defined by the variance of the parameters

$$p(D|f_i) \approx p(D|\mathbf{w}_{MP}, f_i)p(\mathbf{w}_{MP}|f_i)\text{det}^{-\frac{1}{2}}(A/2\pi),$$

where $A = -\nabla^2 \text{ln} p(\mathbf{w}|D, f_i)$ — Hessian at $\mathbf{w}_{MP}$. The variable $\sigma_{w|D}$ depends on the posterior distribution of the parameters $\mathbf{w}$.
The $p(\mathbf{w}_{MP}|f_i) = 1/\sigma_w$ and

$$\text{Occam factor} = \frac{\sigma_{w|D}}{\sigma_w}.$$

The Occam factor shows the «compression» of the parameter space when the data have come.


P(D|f3)
P(D|f2)
P(D|f1)
P(w|D,f1)
σ w|D
P(w|f1)
σ w|D
w MP
w
P(w|D,f2)
P(w|f2)
P(w|D,f3)
P(w|f3)

The indexes of

- objects are $\{1, \ldots, i, \ldots, m\} = \mathcal{I}$, the split $\mathcal{I} = \mathcal{B}_1 \sqcup \cdots \sqcup \mathcal{B}_K$;
- features are $\{1, \ldots, j, \ldots, n\} = \mathcal{J}$, the active set $\mathcal{A} \subseteq \mathcal{J}$.

The regression model

$$f : (\mathbf{w}, \mathbf{x}) \mapsto y;$$

the selected model

$$\mathsf{E}(\mathbf{y}|X) = X_{\mathcal{A}}\mathbf{w}_{\mathcal{A}}, \text{ or } \mathsf{E}(y_i|\mathbf{x}) = \mathbf{w}_{\mathcal{A}}^{\mathsf{T}}\mathbf{x}_i.$$

The multilevel model $\mathfrak{f}$ is a set of the models $\mathfrak{f} = \{f_k | k = 1, \ldots, K\}$, such that for each k

$$\mathsf{E}(y_{i \in \mathcal{B}_k}|\mathbf{x}) = \mathbf{w}_{(k)}^{\mathsf{T}}\mathbf{x}_{i \in \mathcal{B}_k},$$

where

$$\mathcal{I} = \sqcup_{k=1}^{K}\mathcal{B}_k \ni i.$$

Single model:

$$\hat{f}(\mathbf{w}, \mathbf{x}) = \arg \max_{\mathcal{A} \subseteq \mathcal{J}} \mathcal{E}\left(f(\mathbf{w}_{\mathcal{A}}, \mathbf{x})\right).$$

Multilevel model:

$$\hat{\mathfrak{f}}(\mathbf{w}_{(1)}, \dots, \mathbf{w}_{(K)}, \mathbf{x}) = \arg \max_{\sqcup_{k=1}^{K} \mathcal{B}_k = \mathcal{I}} \prod_{k=1}^{K} \mathcal{E}\left(f(\mathbf{w}_{(k)}, \mathbf{x}_{\mathcal{B}_k})\right).$$

Assume the target variable could be approximated by K linear models with parameters $\mathbf{w}_{(k)} \in \mathbb{R}^n$.
Then the distribution of the target variable y for the mixture of normal distributions is

$$p(y|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(y|\mathbf{w}_{(k)}^{\mathsf{T}}\mathbf{x}, \beta).$$

The parameters $\boldsymbol{\theta}$ are concatenated vectors:

$$\boldsymbol{\theta} = [\mathbf{w}_{(1)}, \dots, \mathbf{w}_{(k)}, \boldsymbol{\pi}, \beta]^{\mathsf{T}},$$

where

- $\mathbf{w}_{(1)}, \dots, \mathbf{w}_{(k)}$ are parameters for each of K models,
- $\boldsymbol{\pi} = [\pi_1, \dots, \pi_k]$ is weighs of the models,
- β variance of y, here the covariance matrix $B = \beta I_m$ for $\mathbf{y}$.

The likelihood logarithm function for given data set $D = \{(y_i, \mathbf{x}_i) | i \in \mathcal{I}\} = (\mathbf{y}, X)$ is

$$\ln p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{i \in \mathcal{I}} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(y | \mathbf{w}_{(k)}^{\mathsf{T}} \mathbf{x}_i, \beta) \right).$$

Introduce the matrix

$$Z = \left[\mathbf{z}_1, \dots, \mathbf{z}_m | \mathbf{z} \in \{0, 1\}^K \right].$$

All the components of $\mathbf{z}_i = [z_{i1}, \dots, z_{ik}]$ equal 0 but for k-th: this data sample is generated by k-th model.

The log-likelihood function for joint distribution of $\mathbf{y}, Z$ is

$$\ln p(\mathbf{y}, Z|\boldsymbol{\theta}) = \sum_{i=1}^{m} \sum_{k=1}^{K} z_{ik} \ln \left(\pi_k \mathcal{N}(y_i | \mathbf{w}_{(k)}^{\mathsf{T}} \mathbf{x}_i, \beta) \right).$$

Set initial $\boldsymbol{\theta}^*$ and estimate the vector $\boldsymbol{\theta}$ and the matrix Z iteratively.

E-step: Introduce the matrix $\Gamma = [\gamma_{ik}]$, as expectation that i-th sample is generated by k-th model,

$$\gamma_{ik} = \mathsf{E}(z_{ik}) = p(k|\mathbf{x}_i, \boldsymbol{\theta}^*) = \frac{\pi_k \mathcal{N}(y_i|\mathbf{w}_{(k)}^{\mathsf{T}}\mathbf{x}_i, \beta)}{\sum_{k'=1}^{K} \pi'_k \mathcal{N}(y_i|\mathbf{w}_{(k)}^{\mathsf{T}}\mathbf{x}_i, \beta).}$$

Use $\Gamma = [\gamma_{ik}]$ to define the posterior distribution $p(Z|\mathbf{y}, \boldsymbol{\theta}^*)$ of the likelihood function

$$Q(\boldsymbol{\theta}) = \mathsf{E}_Z(\ln p(\mathbf{y}, Z|\boldsymbol{\theta})) = \sum_{i\in\mathcal{I}}\sum_{k=1}^{K}\gamma_{ik}\left(\ln \pi_k + \ln \mathcal{N}(y_i|\mathbf{w}_{(k)}^{\mathsf{T}}\mathbf{x}_i, \beta)\right).$$

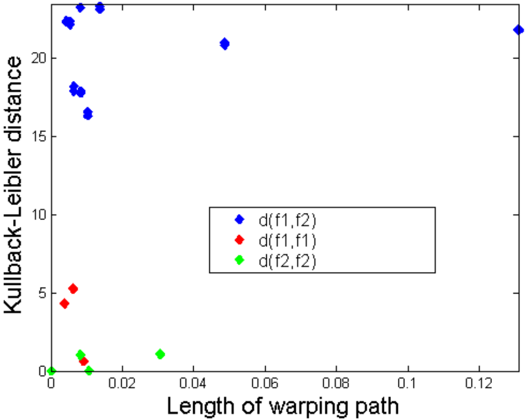
M-step: Maximize function $Q(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, where the matrix Γ is fixed. The model weight coefficients must be normalized, $\sum_{k=1}^{K}\pi_k = 1$.

Introduce a distance function $\rho(f_k, f_l)$ between two models. Use the Jensen-Shannon divergency; $\rho_{kl} \in [0, 1]$ is a metric:

$$\rho(p_k \| p_l) = 2^{-1} D_{\text{KL}}\left(p_k \| p'\right) + 2^{-1} D_{\text{KL}}\left(p' \| p_l\right),$$

where $p' = 2^{-1}(p_k + p_l)$ and $p_k \stackrel{\text{def}}{=} (p(\mathbf{w}|D, A, B, f_k)$. The non-symmetric Kullback-Leibler divergency is

$$D_{\text{KL}}\left(p \| p'\right) = \int\limits_{\mathbf{w} \in \mathcal{W}} p'(\mathbf{w}) \ln \frac{p(\mathbf{w})}{p'(\mathbf{w})} d\mathbf{w}.$$



Fifteen pairs of dots could be separated in the JS metric space (y-axis), but hardly separated in the DTW space (x-axis).

See

mvr.svn.sourceforge.net/viewvc/mvr/lectures/Strijov2012IAM.METU.Part4.pdf

or for short

bit.ly/K3i8zJ