PA196: Pattern Recognition

11. Additional topics

Dr. Vlad Popovici
popovici@iba.muni.cz

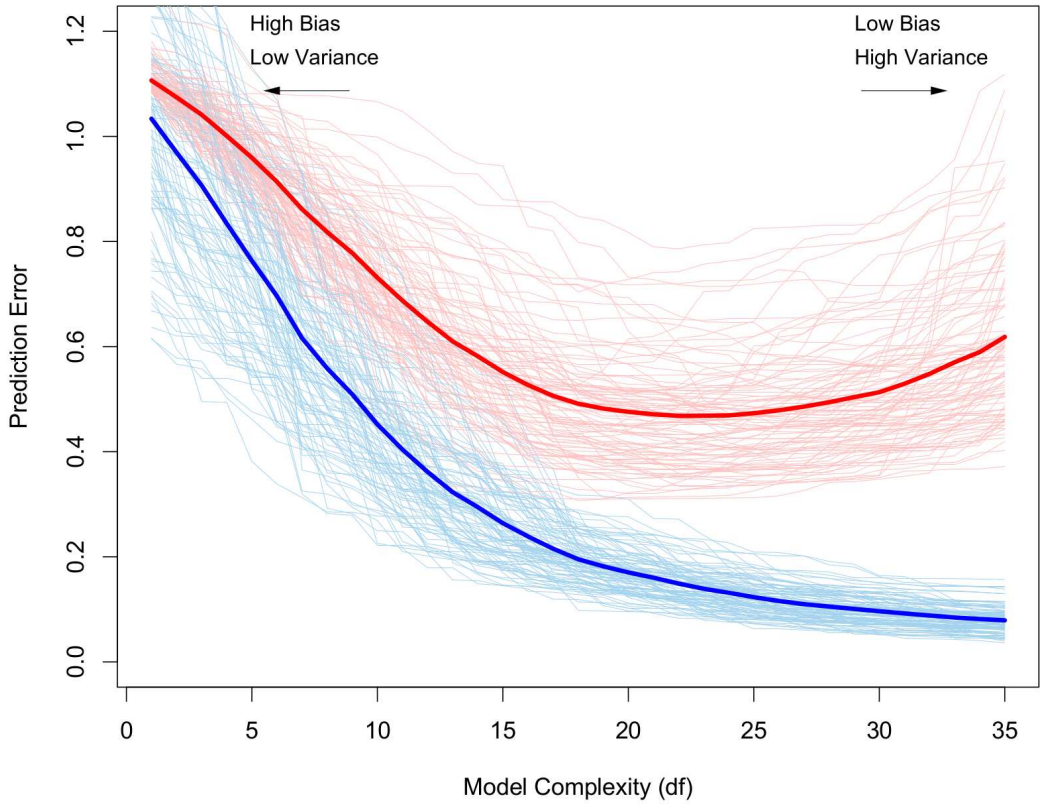
Institute of Biostatistics and Analyses
Masaryk University, Brno

Outline

Outline

Bias-variance decomposition

Hastie et al. Elements of Statistical Learning, fig.7.1



[Penalized regression with various model complexity parameters]

- assume some training set $\mathcal{Z} = \{(\mathbf{x}_i, y_i)|i = 1, \ldots, n\}$ has been drawn i.i.d. from the underlying fixed distribution $P(X, Y)$
- let $L(y, h(\mathbf{x}))$ be the loss function (h is the classifier function)
- training error (apparent error):

$$\text{Err}_0 = \frac{1}{n} \sum_{i=1}^{n} L(y_i, h(\mathbf{x}_i))$$

- generalization error: the prediction error over a test set:

$$\text{Err}_{\mathcal{Z}} = \text{E}[L(Y, h(X))|\mathcal{Z}]$$

- expected prediction error (expected loss):

$$\text{Err} = \text{E}[\text{Err}_{\mathcal{Z}}] = \text{E}[L(Y, h(X))]$$

Loss functions (some):

- 0-1 loss: $L(y, h(\mathbf{x})) = \mathbf{1}_{y \neq h(\mathbf{x})}$
- squared loss: $L(y, h(\mathbf{x})) = (y - h(\mathbf{x}))^2$
- in the context of MAP: for K groups/classes, $1, \ldots, K$ $p_k(X) = \Pr(Y = k|X)$ and the classifier is (a monotone transformation of) the estimate $\hat{p}_h$. Then an adequate loss is

$$\begin{aligned} L(Y, \hat{p}_k) &= -2 \sum_{k=1}^{K} \mathbf{1}_{Y=k} \log \hat{p}_k(X) \\ &= -2 \log \hat{p}_Y(X) \\ &= -2 \times \text{log-likelihood} \end{aligned}$$

("-2" is used to make above loss equivalent to squared loss under Gaussian distributions)

Under some model $Y = h(X) + \epsilon$, where ϵ is some noise ($\mathrm{E}[\epsilon] = 0$, $\mathrm{Var}[\epsilon] = \sigma_\epsilon^2$, and using squared loss, the error at a given point $\mathbf{x}_0$ can be written as

$$\begin{aligned}\mathrm{Err}(\mathbf{x}_0) &= \mathrm{E}[(Y - h(X))^2 | X = \mathbf{x}_0] \\ &= \sigma_\epsilon^2 + [\mathrm{E}[h(\mathbf{x}_0)] - Y]^2 + \mathrm{E}[h(\mathbf{x}_0) - \mathrm{E}[h(\mathbf{x}_0)]]^2 \\ &= \sigma_\epsilon^2 + \mathrm{Bias}^2 + \mathrm{Variance}\end{aligned}$$

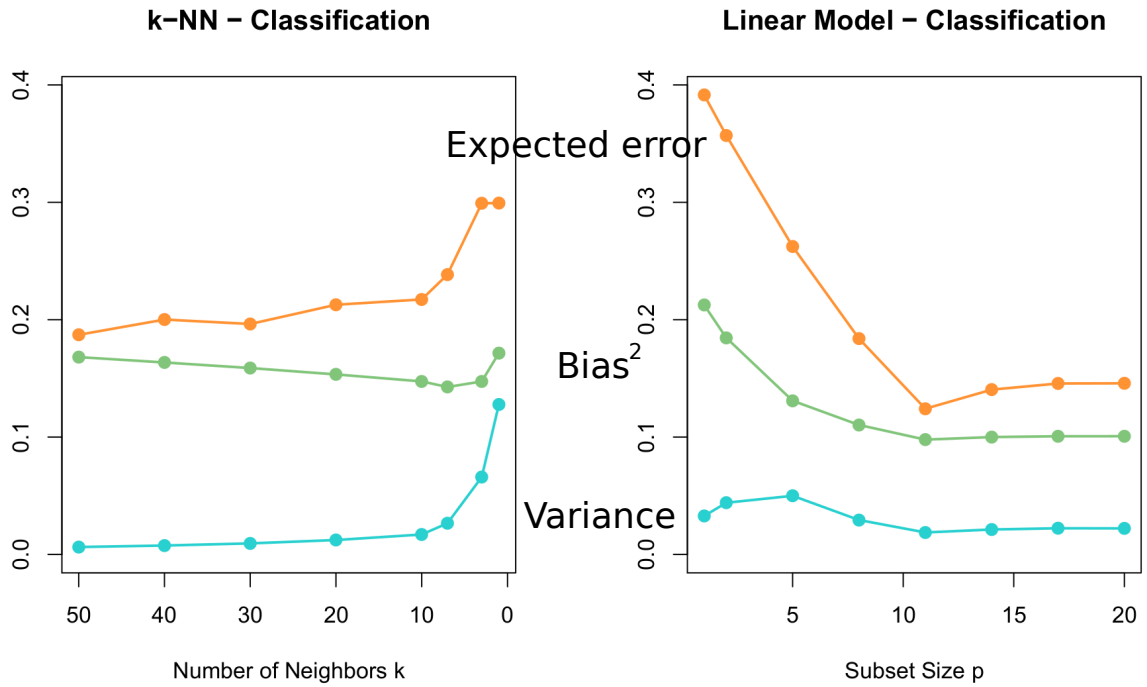
- σ_ϵ^2 cannot be influenced by the model
- the bias: difference between true value and predicted value
- the variance: the expected squared deviation of prediction from its mean
- too complex models: low bias, high variance
- too simple models: high bias, low variance

Model complexity

- in some cases, it is easy to quantify the model complexity
- k–NN: $1/k$ is a measure of complexity
- for a linear model $h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$, the complexity is directly related to the number of non-zero coefficients
- SVM: VC-dimension can be interpreted as a measure of complexity
- "Occam's razor" principle (*lex parsimoniae*): among competing hypotheses/explanations the "simpler" one (with fewest assumptions) should be preferred

Example:

Hastie et al. Elements of Statistical Learning, fig.7.2



Outline

- idea: compute some fitness indicator for a series of models and choose the "best" one
- for classifiers, the most used fitness indicator is the classification performance → estimate the error rate or AUC or any other performance parameter, for a series of values of meta-parameter(s) and choose the one with lowest error rate (or highest AUC, etc). E.g.: grid search we used for SVM
- alternative: try to balance the model complexity and its fitness: AIC, BIC, MDL

AIC - Akaike's Information Criterion

In general, for log-likelihood maximization criterion for model fitting,

$$\text{AIC} = -\frac{2}{n} \times \text{log-likelihood} + 2\frac{d}{n}$$

where "log-likelihood" is the fitted (maximized) log-likehood (by the classifier) and d is a measure of complexity (degrees of freedom) of the model.
The best model (in AiC-sense): the one that minimizes AIC.

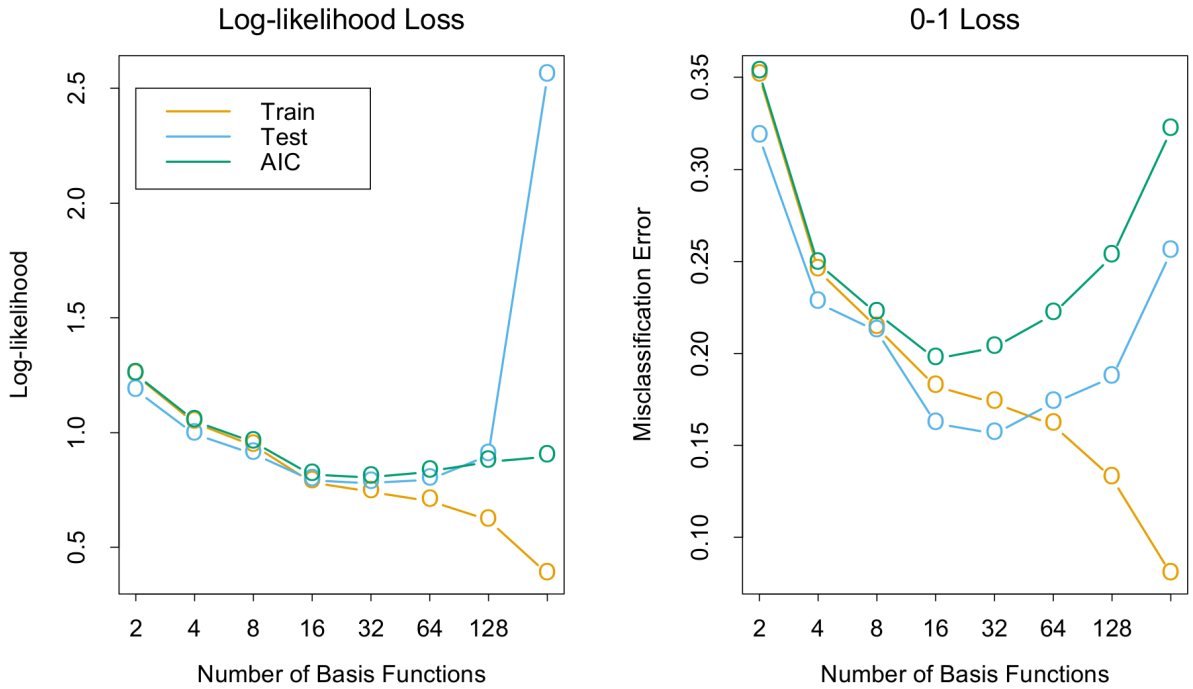
AIC - for classifiers

$$\text{AIC}(\alpha) = \text{Err}_0(\alpha) + 2\frac{d(\alpha)}{n}\hat{\sigma}_\epsilon^2$$

- α is some meta-parameter of the classifier (e.g. polynomial degree for a polynomial kernel SVM, or k in $k-$NN etc.)
- $d(\alpha)$ is the corresponding complexity
- AIC(α) is an estimate of the test error curve
- best model (in AIC-sense) is the one with α minimizing AIC(α)
- $\hat{\sigma}_\epsilon^2$ can be estimated from mean squared error of a low-bias model

Example

Hastie et al. Elements of Statistical Learning, fig.7.3



BIC - Bayesian Information Criterion

In general, for log-likelihood-maximization settings,

$$\text{BIC} = -2 \times \text{log-likelihood} + d \log n$$

which, for a classifier, can be written as

$$\text{BIC} = \frac{n}{\hat{\sigma}_\epsilon^2}\left[\text{Err}_0 + \log n \cdot \frac{d}{n}\hat{\sigma}_\epsilon^2\right]$$

- BIC penalizes more heavily complex models (than AIC)
- the best model (in BIC sense) is the one that minimizes BIC

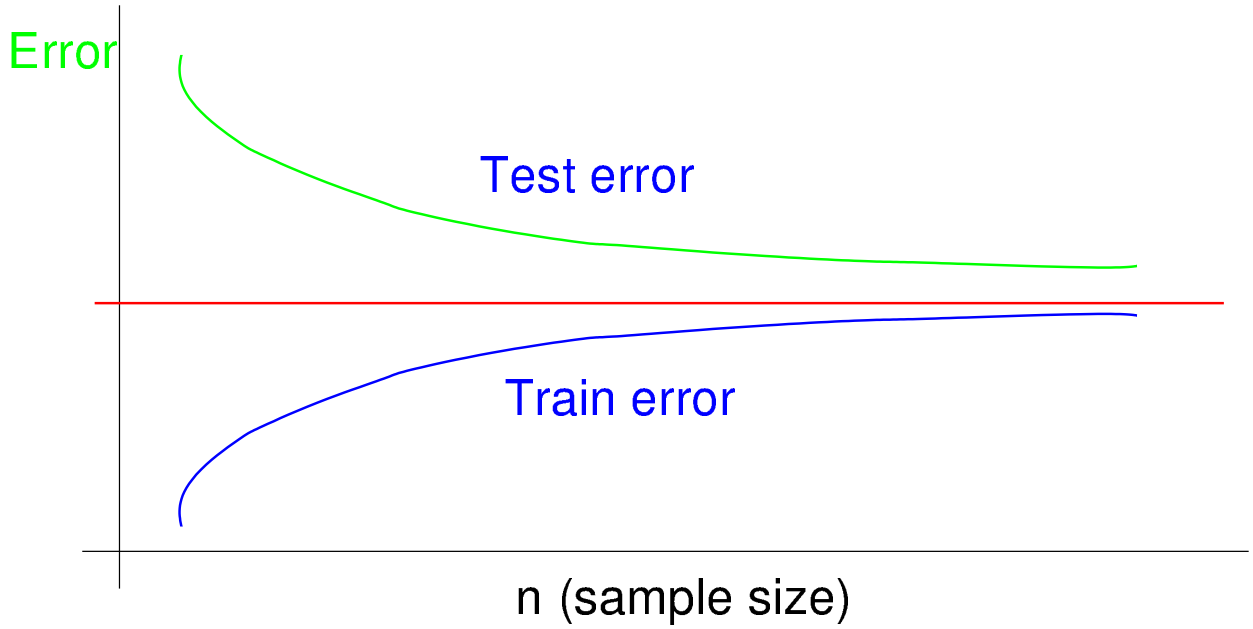
MDL - Minimum Description Length

- MDL leads to a formally identical criterion to BIC, but comes from a totally different theoretical framework
- the classifier is seen as an *encoder* of the *message* (data)
- the model is the encoded message to be transmitted - hence we want it to be *parsimonious* (sparse) and with limited information loss

Outline

Learning curves

- a "diagnostic" for classifier training
- can be used to estimate/approximate the sample size needed for a given problem



Example: Popovici et al, Effect of training-sample size..., BCR 2010

- breast cancer gene expression data
- problems: prediction of ER status, pCR and pCR within ER-
- the performance (AUC) is estimated for increasing sample size
- the following learning curve model is fit (Fukunaga):

$$AUC(n) = a + b/n$$

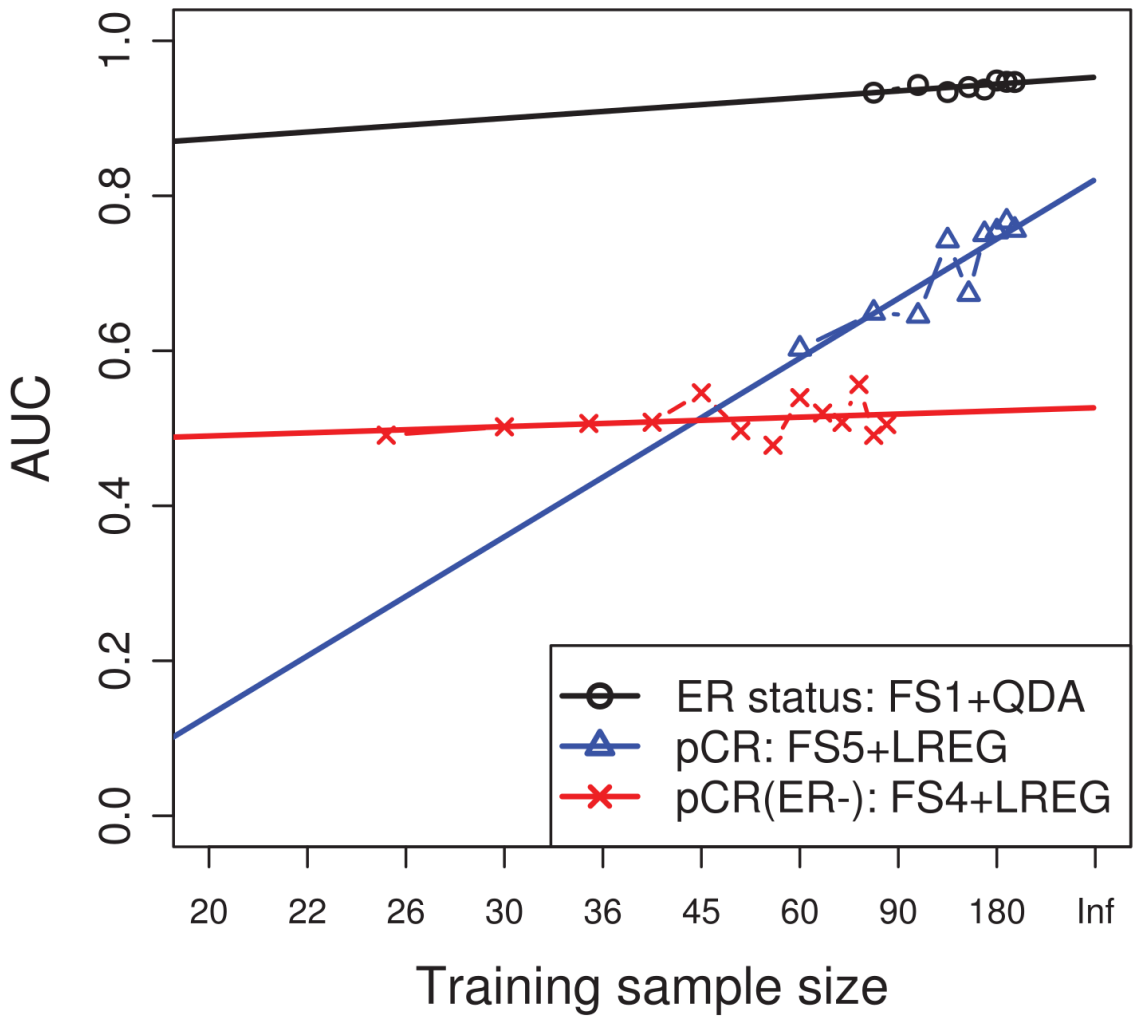


Figure 4 Learning curves for the best predictors for each of the three endpoints. For each endpoint, the learning curve of the best-performing model on the validation set was estimated by fivefold cross-validation for gradually increasing sample sizes. The plot shows both the estimated performance for different sample sizes and the fitted curve. The quadratic discriminant analysis (QDA) classifier required more than 60 samples, so the minimum sample size for it was 80. Note the nonlinear scale of the x-axis.