STK2100- Machine Learning and Statistical Methods for Prediction and Classification

Spring 2026

Model assessment and selection

Covers Sections 7.1-7.7 and 7.10-7.11

Ingrid Hobæk Haff
Department of Mathematics
University of Oslo

Evaluate and select models

- Assume that we have several possible models
- How do we decide which the best?
- How do we evaluate the chosen model?
- This should be done based on measures of performance on new data
- Model selection: estimate the performance of different models in order to choose the best one
- Model assessment: having chosen a final model, estimate its prediction error (generalisation error) on new data

Evaluate and select models

- One strategy to do both model selection and model assessment is to divide the full dataset randomly into three disjoint datasets:
  - a training set used to estimate/train the model
  - a validation set used to select model
  - a test set used to evaluate the final model

- However, this requires large datasets
- We will look at two alternatives:
  - information criteria (AIC, BIC, Mallows C_p) to analytically approximate the validation or test step
  - efficient re-use of the data for model selection or evalutation through cross-validation or the bootstrap

Regression and the prediction error under squared error loss

- Regression: $Y = f(\mathbf{x}) + \varepsilon$, with $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$
- The prediction $\widehat{Y} = \widehat{f}(\mathbf{x})$ of Y for a new $\mathbf{x}$ is based on observed data $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$
- Note that $Y = f(\mathbf{x}) + \varepsilon$ is independent of $\widehat{Y}$, so that ε is independent of $\widehat{Y}$
- Using the squared error loss, we have seen that the prediction error decomposes to

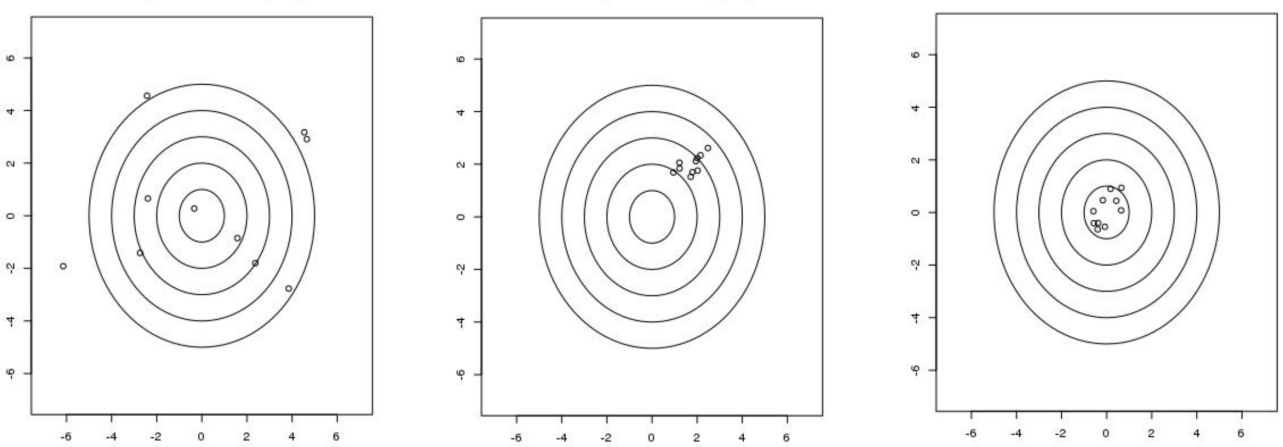
$$\begin{aligned} \mathrm{E}\left(\left(\widehat{Y} - Y\right)^2\right) &= \mathrm{E}\left(\left(\widehat{f}(\mathbf{x}) - f(\mathbf{x})\right)^2\right) + \sigma_\varepsilon^2 \\ &= \text{Reducible error} + \text{Irreducible error} \end{aligned}$$

- Evaluation of the model requires an estimate of the prediction error

Regression and the mean squared error

Further, we have seen that the reducible part of the prediction error can be decomposed as

$$\begin{aligned} \mathrm{E}\left(\left(\widehat{f}(\mathbf{x}) - f(\mathbf{x})\right)^2\right) &= \mathrm{Var}\left(\widehat{f}(\mathbf{x})\right) + \left(\mathrm{E}\left(\widehat{f}(\mathbf{x})\right) - f(\mathbf{x})\right)^2 \\ &= \text{Variance} + \text{Bias}^2 \end{aligned}$$



Large variance, small bias

Small variance, large bias

Small variance and bias

"Average" over many $\mathbf{x}$'s

- Note that $\mathrm{E}\left(\left(\widehat{f}(\mathbf{x}) - Y(\mathbf{x})\right)^2\right)$ is only related to a particular $\mathbf{x}$
- In practice, we want to do predictions on many future $\mathbf{x}$s
- We do not know these, but assume they are drawn from some distribution $p(\mathbf{x})$
- It is more natural to look at the expected prediction error

$$\mathrm{Err} = \int_{\mathbf{x}} \mathrm{E}\left(\left(\widehat{f}(\mathbf{x}) - Y(\mathbf{x})\right)^2\right) p(\mathbf{x}) d\mathbf{x} = \mathrm{E}_{\mathbf{X}}\left(\mathrm{E}_{\mathbf{Y}|\mathbf{X}}\left(\left(\widehat{f}(\mathbf{X}) - Y(\mathbf{X})\right)^2 | \mathbf{X}\right)\right)$$

- Sometimes we are more interested in the generalisation error

$$\mathrm{Err}_{\mathscr{T}} = \mathrm{E}_{\mathbf{X}}\left(\mathrm{E}_{\mathbf{Y}|\mathbf{X}}\left(\left(\widehat{f}(\mathbf{X}) - Y(\mathbf{X})\right)^2 | \mathbf{X}\right) | \mathscr{T}\right),$$

  where $\mathscr{T} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$ is the training data
- The expected prediction error Err is then an "average" over $\mathrm{Err}_{\mathscr{T}}$

$$\mathrm{Err} = \mathrm{E}_{\mathscr{T}}\left(\mathrm{Err}_{\mathscr{T}}\right)$$

- $\mathrm{Err}_{\mathscr{T}}$ is typically of interest, but Err is easier to analyse theoretically

General loss functions

- Typical loss functions in regression are

$$L(Y, \widehat{f}(\mathbf{X})) = \begin{cases} (Y - \widehat{f}(\mathbf{X}))^2 & \text{squared error} \\ |Y - \widehat{f}(\mathbf{X})| & \text{absolute error} \end{cases}$$

- Typical loss functions in classification are

$$L(G, \hat{G}(\mathbf{X})) = \begin{cases} I(G \neq \hat{G}(\mathbf{X})) & \text{0-1 loss} \\ -2 \log \hat{p}_G(\mathbf{X}) & -2 \times \text{log-lik} \end{cases}$$

with $\log \hat{p}_G(\mathbf{X}) = \sum_{k=1}^{K} I(G = k) \log \hat{p}_k(\mathbf{X})$

- The generalisation error is usually of interest:

$$\text{Err}_{\mathscr{T}} = \begin{cases} \mathrm{E}_{\mathbf{X}} \left(\mathrm{E}_{\mathbf{Y}|\mathbf{X}} \left(L(Y, \widehat{f}(\mathbf{X})) | \mathbf{X} \right) | \mathscr{T} \right) & \text{Regression} \\ \mathrm{E}_{\mathbf{X}} \left(\mathrm{E}_{\mathbf{G}|\mathbf{X}} \left(L(G, \widehat{f}(\mathbf{X})) | \mathbf{X} \right) | \mathscr{T} \right) & \text{Classification} \end{cases}$$

- The expected generalisation error is then given by

$$\text{Err} = \begin{cases} \mathrm{E}_{\mathbf{X}} \left(\mathrm{E}_{\mathbf{Y}|\mathbf{X}} \left(L(Y, \widehat{f}(\mathbf{X})) | \mathbf{X} \right) \right) & \text{Regression} \\ \mathrm{E}_{\mathbf{X}} \left(\mathrm{E}_{\mathbf{G}|\mathbf{X}} \left(L(G, \widehat{f}(\mathbf{X})) | \mathbf{X} \right) \right) & \text{Classification} \end{cases}$$

Estimation of the expected generalisation error

- The distribution $p(\mathbf{x})$ of $\mathbf{X}$ is unknown, but we can approximate Err by the in-sample error
- Assume that $Y_1^0, \ldots, Y_N^0$ are new responses corresponding to the observed covariate values $\mathbf{x}_1, \ldots, \mathbf{x}_N$
- The in-sample error is then given by

$$\mathrm{Err}_{in} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}_{Y^0} \left(L(Y_i^0, \widehat{f}(\mathbf{x}_i)) | \mathscr{T} \right) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{Err}_{in}(\mathbf{x}_i)$$

- This may again be estimated by the training error

$$\overline{\mathrm{err}} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \widehat{f}(\mathbf{x}_i))$$

Optimism of the training error rate

- $Y(\mathbf{x})$ is independent of $\widehat{f}(\mathbf{x})$, as $\widehat{f}(\mathbf{x})$ is based on $\{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$, and $(Y, \mathbf{x})$ are indepdendent new data
- However, Y_i and $\widehat{f}(\mathbf{x}_i)$ are dependent
- This leads to optimism, as the expected generalisation error is underestimated by $\overline{\text{err}}$, i.e.

$$\overline{\text{err}} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \widehat{f}(\mathbf{x}_i)) \overset{typically}{<} \text{Err}_{\mathscr{T}}$$

- The optimism is then defined as

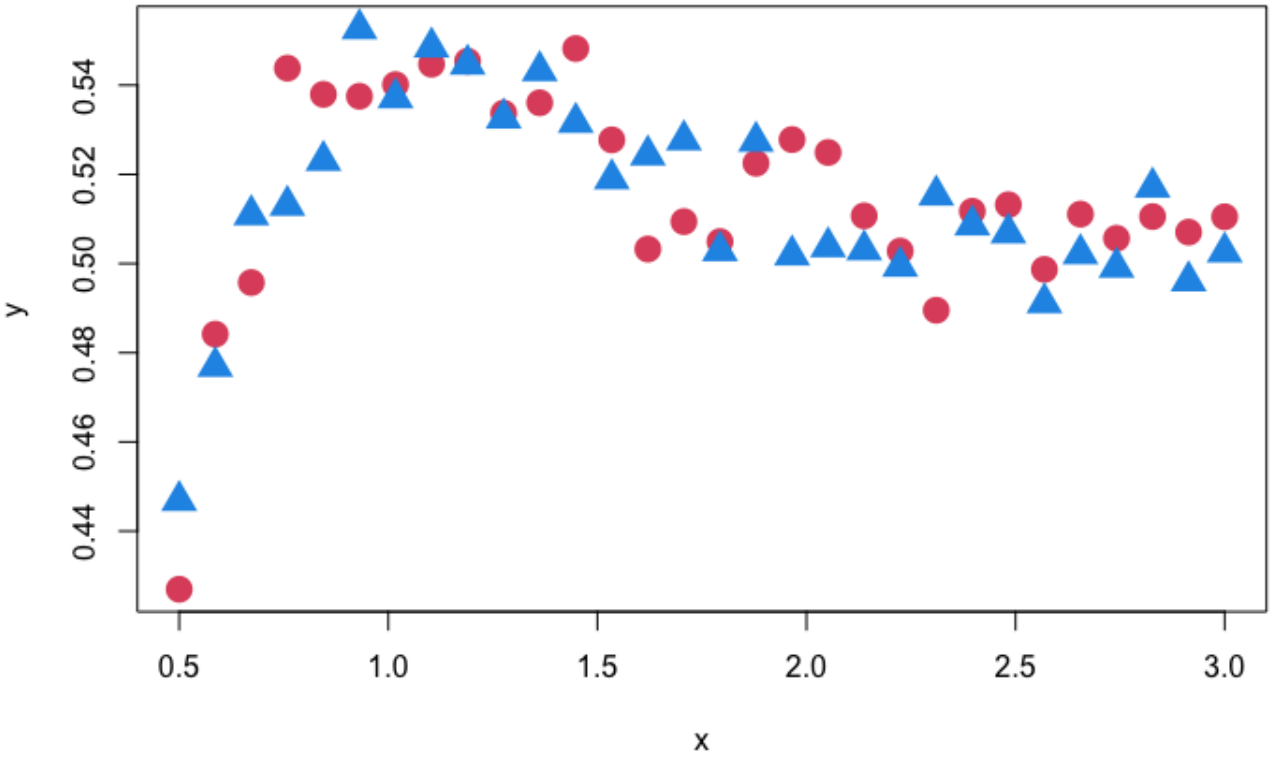
$$\text{op} \equiv \text{Err}_{in} - \overline{\text{err}}$$

and the expected optimism as

$$\omega \equiv \text{E}_{\mathscr{T}}(\text{op})$$

Example: Yesterday data

Training data (x_i, y_i), $i = 1, \ldots, 30$ (red dots), as well as an independent test set of the same size (blue triangles), generated from the model $Y = f(x) + \varepsilon$, with $\varepsilon \sim N(0, 0.01^2)$ and some $f(x)$



See R code

The Bias-Variance Decomposition

- Assume that $Y = f(\mathbf{x}) + \varepsilon$, with $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$
- We have seen that for the squared error loss

$$\begin{aligned}
\mathrm{Err}(\mathbf{x}^0) &= \mathrm{E}\left(\left(Y - \widehat{f}(\mathbf{X})\right)^2 | \mathbf{X} = \mathbf{x}^0\right) \\
&= \sigma_\varepsilon^2 + \mathrm{E}\left(\left(f(\mathbf{X}) - \widehat{f}(\mathbf{X})\right)^2 | \mathbf{X} = \mathbf{x}^0\right) \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}(\widehat{f}(\mathbf{x}^0))^2 + \mathrm{Var}(\widehat{f}(\mathbf{x}^0)) \\
&= \text{Irreducible error} + \mathrm{Bias}^2 + \mathrm{Variance}
\end{aligned}$$

- The irreducible error is unavoidale, whereas the bias and variance depend on the chosen estimator $\widehat{f}$ of f
- For the k-nearest neighbour regression, with $\widehat{f}(\mathbf{x}^0) = \frac{1}{k}\sum_{l=1}^{k} f(\mathbf{x}_{(l)})$, we obtain

$$\mathrm{Err}(\mathbf{x}^0) = \sigma_\varepsilon^2 + \left(f(\mathbf{x}^0) - \frac{1}{k}\sum_{l=1}^{k} f(\mathbf{x}_{(l)})\right)^2 + \frac{1}{k}\sigma_\varepsilon^2,$$

where the $\mathbf{x}_{(l)}$s are the k nearest neighbours to $\mathbf{x}^0$

Linear regression

- In linear regression, we have $\widehat{f}(\mathbf{x}) = \mathbf{x}^T\hat{\beta}$, where $\hat{\beta}$ (typically) is obtained by least squares estimation
- This gives

$$\text{Err}(\mathbf{x}^0) = \sigma_\varepsilon^2 + \underbrace{\left(f(\mathbf{x}^0) - \text{E}\left(\widehat{f}(\mathbf{x}^0)\right)\right)^2}_{Bias^2} + \underbrace{||\mathbf{h}(\mathbf{x}^0)||^2\sigma_\varepsilon^2}_{Variance},$$

  where $\mathbf{h}(\mathbf{x}^0) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^0$
- Further, it may be shown that $\frac{1}{N}\sum_{i=1}^N ||\mathbf{h}(\mathbf{x}_i)||^2 = \frac{d}{N}$, where d is the number of covariates (including the intercept)
- This gives

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \text{Err}(\mathbf{x}_i) &= \frac{1}{N}\sum_{i=1}^N \left(\sigma_\varepsilon^2 + \left(f(\mathbf{x}_i) - \text{E}\left(\widehat{f}(\mathbf{x}_i)\right)\right)^2 + ||\mathbf{h}(\mathbf{x}_i)||^2\sigma_\varepsilon^2\right) \\ &= \sigma_\varepsilon^2 + \frac{1}{N}\sum_{i=1}^N \left(f(\mathbf{x}_i) - \text{E}\left(\widehat{f}(\mathbf{x}_i)\right)\right)^2 + \frac{d}{N}\sigma_\varepsilon^2\end{aligned}$$

Expressions for the optimism

- It may be shown show that for any prediction method with quadratic loss:

$$\omega = \mathrm{E}_{\mathscr{T}}(\mathrm{op}) = \frac{2}{N}\sum_{i=1}^{N}\mathrm{Cov}(\hat{Y}_i, Y_i)$$

- This gives

$$\mathrm{E}_{\mathscr{T}}(\mathrm{Err}_{in}) = \mathrm{E}_{\mathscr{T}}(\overline{\mathrm{err}}+\mathrm{op}) = \mathrm{E}_{\mathscr{T}}(\overline{\mathrm{err}})+\mathrm{E}_{\mathscr{T}}(\mathrm{op}) = \mathrm{E}_{\mathscr{T}}(\overline{\mathrm{err}})+\frac{2}{N}\sum_{i=1}^{N}\mathrm{Cov}(\widehat{Y}_i, Y_i)$$

- For linear regression with d covariates and least squares fit, we get

$$\mathrm{E}_{\mathscr{T}}(\mathrm{Err}_{in}) = E_{\mathscr{T}}(\overline{\mathrm{err}}) + \frac{2d}{N}\sigma_{\varepsilon}^2$$

- The above expression is actually valid for any model of the form $Y = f(\mathbf{X}) + \varepsilon$, fitted with quadratic loss, where d is the so-called effective number of parameters

Estimation of the in-sample prediction error

- The in-sample prediction error Err_{in} is typically estimated by
$$\widehat{\text{Err}}_{in} = \overline{\text{err}} + \hat{\omega},$$
where $\hat{\omega}$ is an estimate of the expected optimism
- For additive error models $Y = f(\mathbf{X}) + \varepsilon$ with d effective parameters, this leads to
$$\widehat{\text{Err}}_{in} = \overline{\text{err}} + \frac{2d}{N}\hat{\sigma}_{\varepsilon}^2 = C_p,$$
where $\hat{\sigma}_{\varepsilon}^2$ is an estimate of $\text{Var}(\varepsilon)$ and C_p is Mallow's C_p statistic

The Akaike information criterion (AIC)

- For a model with density/point mass function $p(y|\mathbf{x};\boldsymbol{\theta})$, that is to be fitted with the observations $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, N$, the log-likelihood function is given by

$$l(\boldsymbol{\theta}) = \log\left(\prod_{i=1}^{N} p(y_i|\mathbf{x}_i;\boldsymbol{\theta})\right) = \sum_{i=1}^{N} \log\left(p(y_i|\mathbf{x}_i;\boldsymbol{\theta})\right)$$

- The Akaike information criterion (AIC) is then defined as

$$\text{AIC} = -2l(\hat{\boldsymbol{\theta}}) + 2d, \quad d = |\boldsymbol{\theta}|,$$

  where $\hat{\boldsymbol{\theta}}$ typically is the maximum likelihood estimate of $\boldsymbol{\theta}$
- Under certain regularity conditions, $N \cdot AIC$ is an approximately unbiased estimate of $-2\mathrm{E}\left(\log p(Y|\mathbf{X};\boldsymbol{\theta})\right)$
- Hence, the AIC can be used to estimate Err_{in} when the negative log-likelihood is used as the loss function

The Akaike information criterion (AIC)

- For linear regression with Gaussian error distribution and known variance σ_ε^2, we have

$$
\begin{aligned}
l(\hat{\beta}) &= \log\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}(y_i - \mathbf{x}_i^T\hat{\beta})^2\right)\right) \\
&= -\frac{N}{2}\log(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\hat{\beta})^2,
\end{aligned}
$$

so that

$$
\frac{\sigma_\varepsilon^2}{N}\mathrm{AIC} = \sigma_\varepsilon^2\log(2\pi\sigma_\varepsilon^2) + \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\hat{\beta})^2 + \frac{2}{N}d\sigma_\varepsilon^2 = const. + C_p
$$

- Hence, in this case, the AIC is equivalent to Mallow's C_p
- The AIC can be used for model selection, for instance for variable selection, by choosing the one among a list of candidate models having the smallest AIC
- For nonlinear and other complex models, d must be replaced by some measure of model complexity.

The effective number of parameters

- Assume that the model can be written on the form $\widehat{\mathbf{y}} = \mathbf{S}\mathbf{y}$
- Then, the effective number of parameters is given by

$$\mathrm{df}(\mathbf{S}) = \mathrm{Tr}(\mathbf{S})$$

- It can be shown that if $Y = f(\mathbf{x}) + \varepsilon$, then $\sum_{i=1}^{N} \mathrm{Cov}(\widehat{Y}_i, Y_i) = \mathrm{Tr}(\mathbf{S})\sigma_\varepsilon^2$
- This leads to the more general definition

$$\mathrm{df}(\widehat{Y}) = \frac{\sum_{i=1}^{N} \mathrm{Cov}(\widehat{Y}_i, Y_i)}{\sigma_\varepsilon^2}$$

- For error functions $R(\omega)$ with decay penalty $\alpha \sum_m \omega_m$, we have

$$\mathrm{df}(\alpha) = \sum_{m=1}^{M} \frac{\theta_m}{\theta_m + \alpha}$$

where θ_m are the eigenvalues of $H = \partial^2 R(\omega)/\partial\omega\partial\omega^T$

Information criteria (IC)

- When we have sparse data and/or constrained computational power, information criteria like the AIC can be used
- Information criteria are of the form

$$\text{IC} = -2l(\hat{\boldsymbol{\theta}}) + \text{penalty}(d), \quad d = |\boldsymbol{\theta}|$$

- Different choices of penalty give rise to different information criteria
- The most common alternative to the AIC is the BIC, given by

$$\text{BIC} = -2l(\hat{\boldsymbol{\theta}}) + \log(N)d$$

- The BIC penalizes more complex models harder

Information criteria (IC)

- Information criteria can be used for non-nested models, as opposed to for instance likelihood ratio tests
- Further, they can also be used for dependent data
- They require fewer models to fit than other methods, like cross-validation and the bootstrap
- On the other hand, they rely more heavily on model assumptions

Example: Yesterday data
See R code

Dividing the data

- Recall that a way to evalutate model performance is to divide the full dataset randomly into two disjoint datasets:
  - a training set used to estimate/train the model
  - a test set used to evaluate the model
- The advantages of this are that we have an independent data set for evaluation, and that this is easy to apply and understand
- On the other hand, it reduces the dataset used for training the model
- Further, if the same test set is used to choose between models, the final error measure will be too optimistic, and we should really also set aside a validation set for model selection

Leave-one-out cross-validation (LOOCV)

- We want the training set to be as large as possible, but also need a test set for evaluation
- If the amount of data is not huge, an alternative is cross-validation
- Leave-one-out cross-validation works as follows
  - For $i = 1, \ldots, N$, fit the model on all data except $(y_i, \mathbf{x}_i)$, and use the resulting model f^{-i} to predict y_i by $\widehat{f}^{-i}(\mathbf{x}_i)$
  - Estimate the prediction error by

$$\widehat{\mathrm{Err}}_{LOOCV} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \widehat{f}^{-i}(\mathbf{x}_i))$$

- Strengths
  - Uses almost all data for training, hence gives low bias
  - Uses all data in test sets
- Weaknesses
  - Uses (slightly) different models for each prediction
  - Can give high variance as the training sets are so similar
  - Computer-intensive

K-fold cross-validation

- K-fold cross-validation works as follows:
  - Divide the data randomly into K groups (folds)
  - For $k = 1, \ldots, K$, fit the model to all data, except the $(y_i, \mathbf{x}_i)$s in group k, and use the resulting model $\widehat{f}^{-k}$ to predict the y_is in group k by $\widehat{f}^{-k}(\mathbf{x}_i)$
  - Estimate the prediction error by

$$\widehat{\mathrm{Err}}_{CVK} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \widehat{f}^{-k(i)}(\mathbf{x}_i)),$$

    where $k(i)$ is the group that observation i belongs to
- Strengths
  - Only need to train K models, less computer-intensive than LOOCV
  - All observations still used in test sets
  - Can give lower variance since the fitted models are less similar
- Weaknesses
  - Uses more different models for each prediction, which can give higher bias

Cross-validation

- If the modelling procedure consists of several steps, for instance a pre-selection of a subset of covariates, cross-validation must be applied to the full sequence of steps in order to avoid underestimation of the prediction error
- The choice of K is a trade-off between bias and variance, as higher values, for instance $K = N$ for LOOCV, gives lower bias, whereas lower values, typically 5 or 10, gives lower variance
- It also depends on the computational complexity, which increases with K
- For linear models $\widehat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ and squared error loss, generalised cross-validation gives an approximation to LOOCV:

$$\widehat{\mathrm{Err}}_{GCV} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \widehat{f}(\mathbf{x}_i)}{1 - tr(\mathbf{S})/N}\right)^2$$

The bootstrap

- An alternative to cross-validation is the bootstrap
- The bootstrap works as follows: For $b = 1, \ldots, B$
  - Sample $(y_1^{*b}, \mathbf{x}_1^{*b}), \ldots, (y_N^{*b}, \mathbf{x}_N^{*b})$ from the data with replacement
  - Fit the model to the bootstrap sample, and use the resulting model $\widehat{f}^{*b}$ to predict each y_i (from the original data) with $\widehat{f}^{*b}(\mathbf{x}_i)$
- The straightforward bootstrap estimate of the prediction error is

$$\widehat{\mathrm{Err}}_{boot} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \widehat{f}^{*b}(\mathbf{x}_i))$$

- However, this is not a good estimate of the prediction error
- The reason is that there is a lot of overlap between the bootstrap samples, that act as training sets, and the original data, that acts as the test set

The bootstrap

- More specifically,

$$\begin{aligned}\Pr(\text{obs. } i \in \text{boot. sample } b) &= 1 - \Pr(\text{obs. } i \notin \text{boot. sample } b) \\ &= 1 - \left(1 - \frac{1}{N}\right)^N \approx 1 - e^{-1} \approx 0.632\end{aligned}$$

- To reduce the problem of over-fitting, one may use the leave-one-out bootstrap estimate of prediction error

$$\widehat{\mathrm{Err}}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \widehat{f}^{*b}(\mathbf{x}_i)),$$

where C^{-i} is the set of indices of the bootstrap samples that do not contain observation i

The bootstrap

- However, there is still a training set bias, as the average number of distinct observations in each bootstrap sample is approximately $0.632N$
- This motivates the improved prediction error estimate

$$\widehat{\mathrm{Err}}^{(0.632)} = 0.368\ \overline{\mathrm{err}} + 0.632\ \widehat{\mathrm{Err}}^{(1)}$$

Example: Yesterday data

See R code