IA|BE Data Science Certificate

Module 1 on Foundations of machine learning in actuarial sciences
Machine learning basic concepts

Katrien Antonio
LRisk - KU Leuven and ASE - University of Amsterdam

October 5, 2021

Acknowledgement

Some of the figures in this presentation are taken from *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Some of the figures in this presentation are taken from *The Elements of Statistical Learning: Data mining, Inference and Prediction* (Springer, 2009) with permission from the authors: T. Hastie, R. Tibshirani and J. Friedman.

Some of the figures in this presentation are taken from *Applied Predictive Modeling* (Springer, 2013) with permission from the authors: M. Kuhn and K. Johnson.

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

Data is not labeled in any way

SUPERVISED

Predict a category

Predict a number

CLASSIFICATION

«Divide the socks by color»

REGRESSION

«Divide the ties by length»

UNSUPERVISED

Divide by similarity

Find hidden dependencies

Identify sequences

CLUSTERING

«Split up similar clothing into stacks»

ASSOCIATION

«Find what clothes I often wear together»

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

Taken from Machine learning for everyone. In simple words. With real-world examples. Yes, again.

What is predictive modeling?

- Given a response (or outcome) Y and p different predictors $X_1,\ X_2, \ldots, X_p$, we assume

$$\begin{aligned} Y &= f(X_1, X_2, \ldots, X_p) + \epsilon \\ &= f(\boldsymbol{X}) + \epsilon, \end{aligned}$$

  with f some fixed, but unknown function of $X_1, \ldots, X_p$ (the predictors, independent variables, features or just variables) and ϵ a random error term.

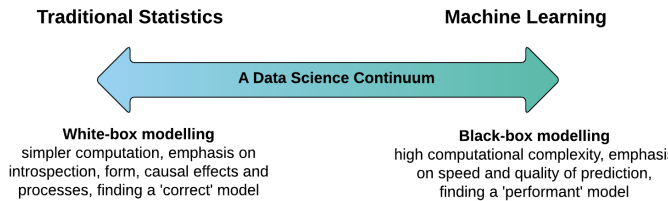
- f represents the systematic information that $\boldsymbol{X}$ provides about Y.

- Predictive modeling refers to a set of approaches for estimating f.

What is predictive modeling?

Why estimate f?

- Prediction
- Inference



What is predictive modeling?

- With prediction: $\hat{Y} = \hat{f}(\boldsymbol{X})$.
- $\hat{f}$ is our estimate for f, often treated as a black box.
- The accuracy of $\hat{Y}$ as a prediction for Y decomposes into (with X and $\hat{f}$ given)

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(\boldsymbol{X}) + \epsilon - \hat{f}(\boldsymbol{X})]^2. \\ &= \underbrace{[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})]^2}_{\text{reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{irreducible}} . \end{aligned}$$

- Reducible error: $\hat{f}$ is not a perfect estimate for f, but potentially the accuracy of $\hat{f}$ can be improved.
- Irreducible error: Y is also a function of ϵ, which can not be predicted using $\boldsymbol{X}$.

What is predictive modeling?

- With inference: how is Y affected as $X_1, \ldots, X_p$ change?
- Thus, we want to understand the relation between $\boldsymbol{X}$ and Y.
- $\hat{f}$ can not be treated as black box; we need its exact form.
- Examples of questions to be answered:
  - which predictors associated with the response?
  - what is the relationship between response and each predictor?
  - can this relationship be summarized using a linear equation, or is it more complicated?

Assessing model accuracy

- *There is no free lunch in statistics!*
- No single method dominates all others over all possible data sets.
- Selecting the best approach can be one of the most challenging parts of performing statistical/machine learning in practice.

Assessing model accuracy

- To evaluate the performance of a statistical/machine learning method on given data:

  evaluate how well predictions actually match observed data.

- Thus, quantify the extent to which the predicted response for a given observation is close to the true response value.

- In regression setting, use e.g. the Mean Squared Error (MSE)

$$\text{MSE} \quad = \quad L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2.$$

Assessing model accuracy

- Examples of (other, more general) loss functions:
  - Residual Sum of Squares (RSS)

$$\text{RSS}(\theta) = L(\theta) \quad = \quad \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2$$

  - the log-probability (or the log-likelihood)

$$L(\theta) \quad = \quad \sum_{i=1}^{n} \log \Pr_\theta(y_i)$$

  - cross-entropy

$$L(\theta) \quad = \quad -\sum_{i=1}^{n} (y_i \cdot \log\left(p_\theta(x_i)\right) + (1 - y_i) \cdot \log\left(1 - p_\theta(x_i)\right)).$$

Assessing model accuracy

- Suppose that we fit our model on

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

  and get $\hat{f}$.

- We are not really interested in whether $\hat{f}(x_i) \approx y_i$ for the training data.
- Our interest:
  - is $\hat{f}(x_0)$ approximately equal to y_0, where (x_0, y_0) is an unseen test observation not used to train the model.
  - Thus, compute the average test error and select a model for which this average test error is small. This is the test MSE.

Assessing model accuracy

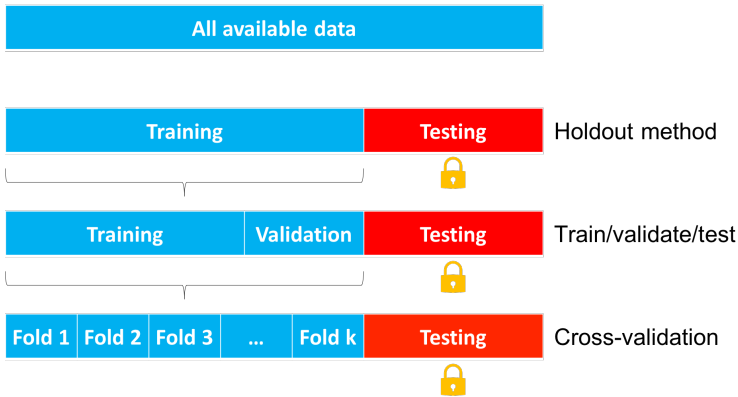
- When a method yields a small training MSE, but a large test MSE

  overfitting results!

- Our statistical learning procedure is working too hard to find patterns in the training data.
- We pick up patterns in the training data caused by random chance. [Signal and the Noise]
- The test MSE will then be very large, because the supposed patterns in the training data are not in the test data.

Assessing model accuracy

Training vs test data



(Picture taken from Introduction to machine learning in R.)

Assessing model accuracy

- U-shape in the test MSE curves is the result of two competing properties of statistical learning methods, see the expected test MSE:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon).$$

- To minimize the expected test error, low variance and low bias are necessary.
- Expected test MSE can never lie below $\text{Var}(\epsilon)$, the irreducible error.

Assessing model accuracy

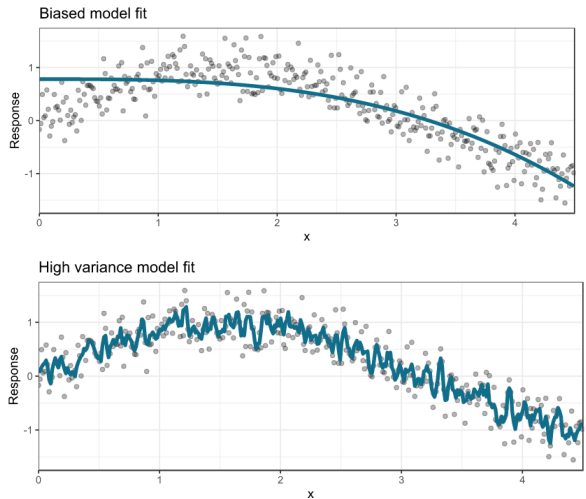
- What do we mean by the variance and bias of a statistical learning method?
- Variance:
  - the amount by which $\hat{f}$ would change if we estimated it using a different training data set
  - using a method with high variance, small changes in the training data can result in large changes in $\hat{f}$
  - more flexible statistical methods have higher variance.

Assessing model accuracy

- What do we mean by the variance and bias of a statistical learning method?
- Bias:
  - the error that is introduced by approximating a real-life problem by a (much simpler) model
  - e.g. if true f is substantially non-linear, no matter how many training observations we have, it will not be possible to produce an accurate estimate using linear regression.

Assessing model accuracy

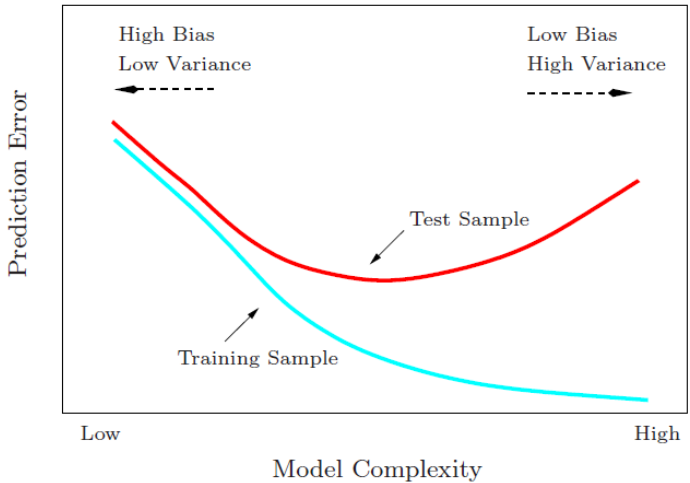
Inspired by Boehmke & Greenwell, 2019, Hands-on machine learning with R, Chapter 2:



Assessing model accuracy

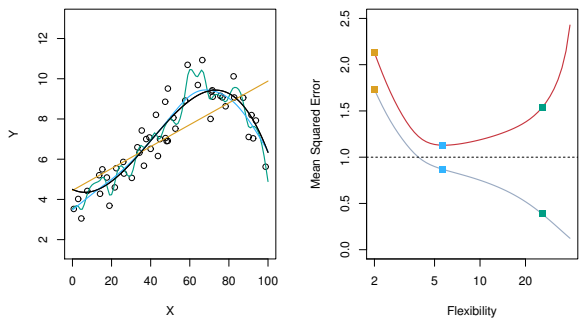
- General rule: with more flexible methods
  - variance will increase and bias will decrease
  - their relative rate of change determines whether the test MSE increases or decreases.

Assessing model accuracy



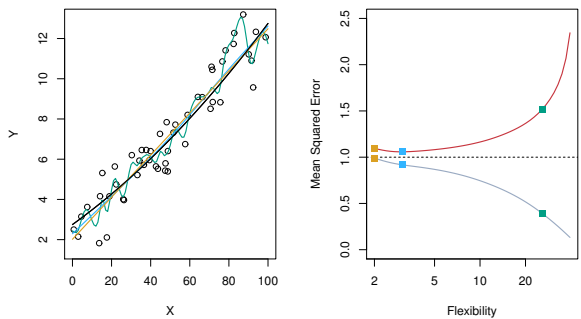
Assessing model accuracy: illustrations

- We generate data from: $Y = f(\boldsymbol{X}) + \epsilon$, with black curve the true f.
- The orange (linear regression), blue (smoothing splines) and green (smoothing splines) curves are three estimates for f, with increasing level of complexity.



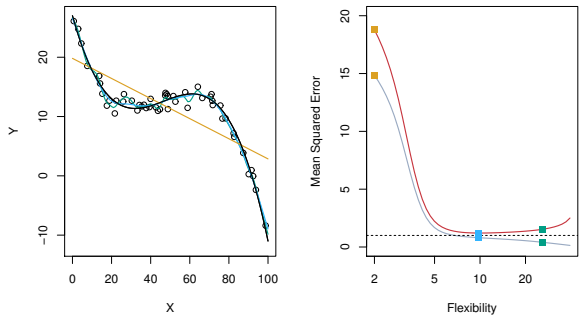
Assessing model accuracy: illustrations

- Another example: true f is now approximately linear.
- Training MSE decreases monotonically as model flexibility increases, and test set has U-shape curve.



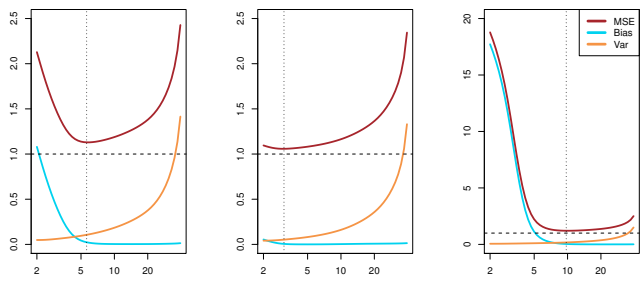
Assessing model accuracy: illustrations

- Last example: true f is highly unlinear.
- Rapid decrease in both curves before test MSE starts to increase slowly.



Assessing model accuracy: illustrations

- We visualize the bias-variance trade off for the three examples considered before.



Assessing model accuracy: classification setting

- Training error rate:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i).$$

- $\hat{y}_i$ is the predicted class label for observation i using $\hat{f}$. $I(.)$ an indicator variable.
- The test error rate associated with a set (x_0, y_0) is

$$\text{Avg}(I(y_0 \neq \hat{y}_0)).$$

- A good classifier is one for which the test error is smallest.

Assessing model accuracy: classification setting

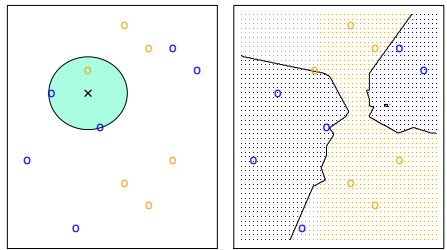
- Example: the K–nearest neighbors (KNN) classifier.
- Given a positive integer K and a test observation x_0, we identify the K points in the training data set that are "closest" to x_0.
- The set $\mathcal{N}_0$ results and we use

$$\Pr(Y = j|X = x_0) \quad = \quad \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

Then, KNN classifies the test observation x_0 to the class with the largest probability.

Assessing model accuracy: classification setting

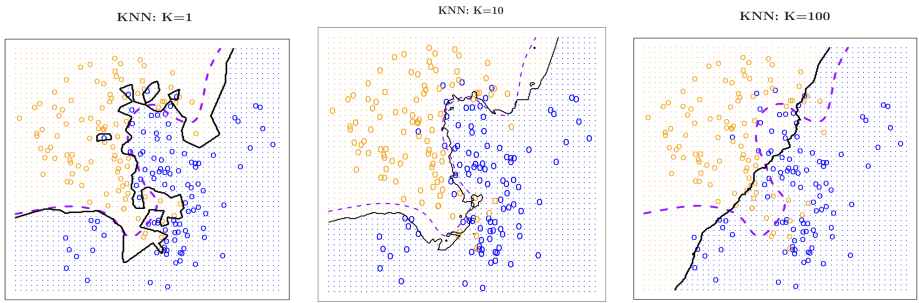
Illustrative example of KNN approach:



(Left) Suppose $K = 3$ and goal is to predict the point labeled by the black cross. (Right) Corresponding KNN decision boundary.

Assessing model accuracy: classification setting

Now compare KNN with K equals 1, 10 and 100.

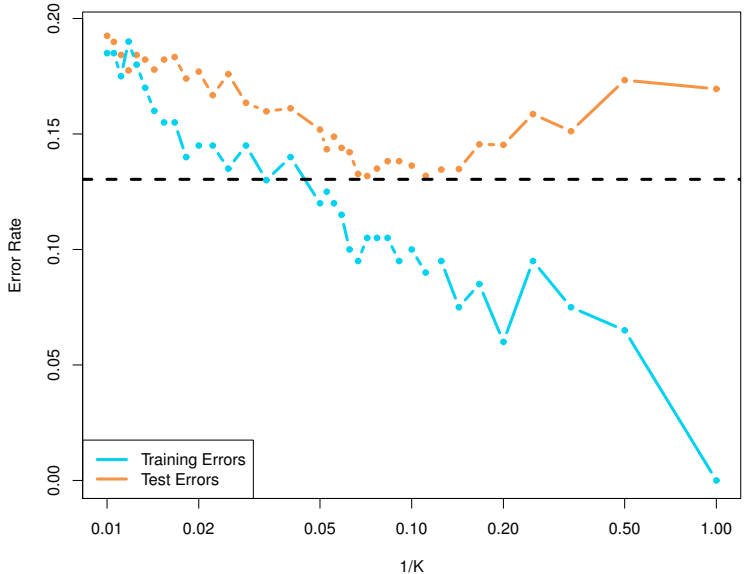


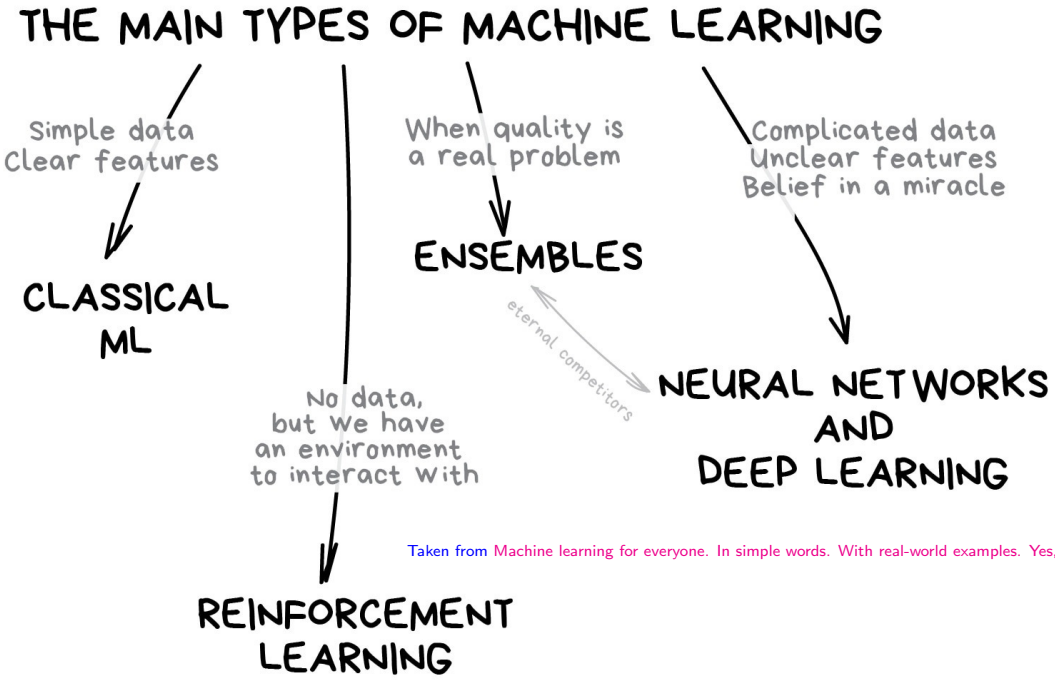
Q: which classifier do you prefer? Which one is over-fitting, under-fitting?

Assessing model accuracy: classification setting

- With $K = 1$, KNN training error rate is 0, but test error rate may be quite high.
- With more flexible classification methods, the training error rate will decline, but the test error rate may not.
- See the plot on the next sheet, where training and test errors are plotted as a function of $1/K$.
- In both regression and classification settings: (model tuning!)
  - choosing correct level of flexibility is critical
  - the bias-variance tradeoff and the resulting U-shape in the test error can make this a difficult task.

Assessing model accuracy: classification setting





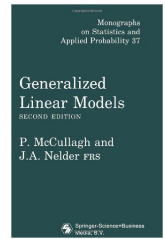
Taken from Machine learning for everyone. In simple words. With real-world examples. Yes, again.

Predictive modeling: a brief history

- Start of the 19th century:
  - work by Legendre and Gauss on method of least squares
  - earliest form of linear regression, focus on quantitative values.
- For qualitative values:
  - Fisher proposed linear discriminant analysis in 1936
  - in 1940s: logistic regression.
- Early 1970s: Generalized Linear Models by Nelder and Wedderburn.

Predictive modeling: a brief history

- Early 1970s: Generalized Linear Models by Nelder and Wedderburn.
  - up to that time: focus almost exclusively on linear methods;
  - fitting other relationships computationally infeasible at that time.

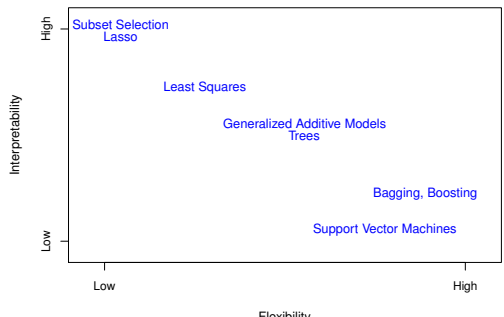


McCullagh & Nelder, 1989

- Mid 1980s: Breiman, Friedman, Olshen and Stonde introduced classification and regression trees.
- Hastie and Tibshirani launched Generalized Additive Models in 1986.
- Since that time, inspired by advent of machine learning, statistical learning emerged as a subfield in statistics.

Predictive modeling: techniques

Trade off between prediction accuracy and model interpretability!



(Taken from James et al., 2013, An introduction to statistical learning.)

Over-fitting

- Modern classification and regression models are highly adaptable:
  - easily overemphasize patterns that are not reproducible
  - model we build should predict new samples with a similar degree of accuracy on the set of data for which the model was evaluated.
- Almost all predictive modeling techniques have tuning parameters that enable the model to flex to find the structure in the data.
- Model tuning:
  - use the existing data to identify settings for the model's parameters
  - use the settings that yield the best and most realistic predictive performance.

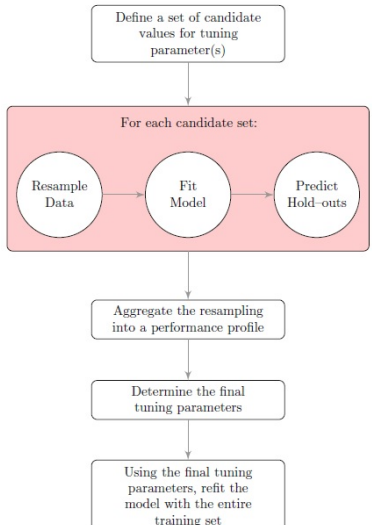
Over-fitting: model tuning

- No analytical formula exists to calculate an appropriate value for such tuning parameters.
- Examples of tuning parameters:
  - K in the K-Nearest Neighbour (KNN) classification model
  - depth of a regression tree
  - cost parameter in a support vector machine
  - ...

Over-fitting: model tuning

- General approach for searching for the best parameters:
  - define a set of candidate values (a grid)
  - generate reliable estimates of model utility across the candidates
  - choose the optimal settings.
- A flowchart of this process is on the next sheet.

Over-fitting: model tuning - flowchart



Over-fitting: model tuning

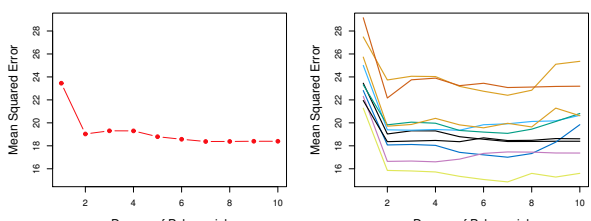
- We need trustworthy estimates of model performance.
- We test the model on samples that were not used for training.
- On these samples we evaluate the set of candidate models (defined by the tuning parameters):
  - possible approach:

    evaluate the models on a test set, but size of the test set may need to be large
  - alternative:

    resample the training set and evaluate the candidate models on these modified versions of the training set.

Cross-validation: validation set approach

- First approach: the validation set approach.
- The validation estimate of the test error rate can be highly variable.
- Example: in a linear regression model explaining `mpg` as a polynomial function of `horsepower`.



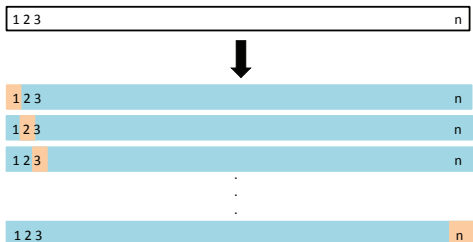
Cross-validation: leave-one-out cross-validation

- Second approach: leave-one-out cross-validation (LOOCV).
- LOOCV splits the set of observations into two parts.
- However, a single observation (x_1, y_1) is used for validation and $\{(x_2, y_2), \ldots, (x_n, y_n)\}$ make up the training set.
- $\text{MSE}_1 = (y_1 - \hat{y}_1)^2$ is an unbiased but highly variable estimate for the test error.
- Repeat this procedure to obtain $\text{MSE}_1, \ldots, \text{MSE}_n$ and calculate the LOOCV estimate for the test MSE as:

$$\text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \text{MSE}_i.$$

Cross-validation: leave-one-out cross-validation

- Second approach: leave-one-out cross-validation (LOOCV).
- LOOCV has less bias than validation set approach. However, LOOCV is potentially very expensive to implement.



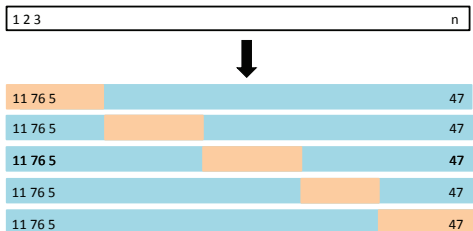
Cross-validation: k-fold cross validation

- Third approach: k-fold cross validation.
- We randomly divide the set of observations into k groups, or folds, of approximately equal size.
- The first fold is treated as a validation set, and the method is fit on the remaining $k-1$ folds. We compute MSE_1 on the observations in the held-out fold.
- We repeat this procedure k times and each time treat a different group of observations as a validation set.
- The k-fold CV estimate is then

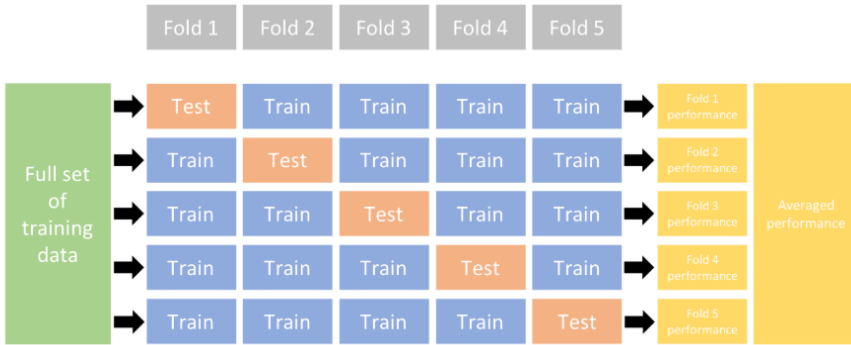
$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_i.$$

Cross-validation: k-fold cross validation

- Third approach: k-fold cross validation.
- In practice: typically use $k = 5$ or $k = 10$ (for computational reasons).



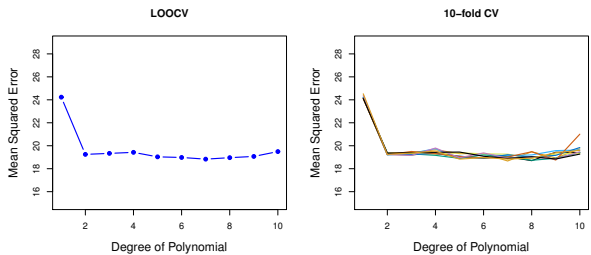
Cross-validation: k-fold cross validation



(Picture taken from Boehmke & Greenwell (2019). Hands-on machine learning with R.)

Cross-validation: k-fold cross validation

- Third approach: k-fold cross validation.
- Example: in the linear regression model explaining `mpg` as a polynomial function of `horsepower`.



Cross-validation: stratified k-fold cross validation

- Fourth approach: stratified k-fold cross validation.
- Let $\mathcal{D}_1, \ldots, \mathcal{D}_k$ be the k folds, i.e. disjoint random subsets of approximately the same size.
- Outliers may fall into the same fold $\mathcal{D}_k$ and this substantially distorts k-fold cross-validation!
- Stratified k-fold cross-validation aims a more equal distribution of outliers across the folds.

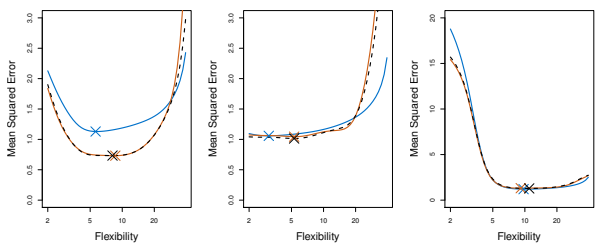
Cross-validation: stratified k-fold cross validation

- Fourth approach: stratified k-fold cross validation.
- How does it work?
  - order the outcomes $Y_{(1)} \geq Y_{(2)} \geq \ldots \geq Y_{(n)}$ (with a deterministic rule in case of ties)
  - build folds $\mathcal{U}_\ell$ (with $\ell = 1, \ldots, \lceil n/k \rceil$) such that $\mathcal{U}_1$ contains k largest observations, $\mathcal{U}_2$ the next k largest observations and so forth
  - construct the folds $\mathcal{D}_k$ as follows

$$\mathcal{D}_k = \{\text{pick randomly from each urn } \mathcal{U}_1, \ldots, \mathcal{U}_{\lceil n/k \rceil} \text{ one case (without replacement)}\}.$$

Cross-validation

- In the examples discussed earlier (cfr. Model Accuracy) the data were simulated, thus: we know the true test MSE.
- We plot: true test MSE (in blue), LOOCV (in black) and 10-fold CV (in orange).
- Interest lies in the location of the minimum point in the estimated test MSE curve.



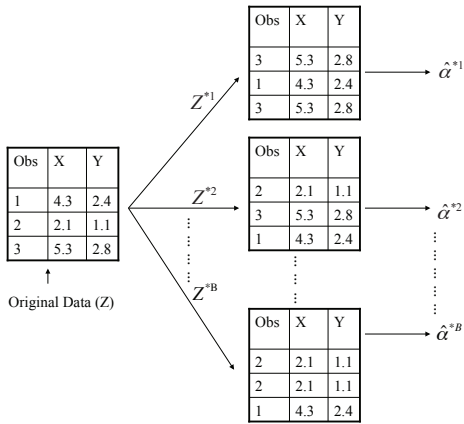
Bootstrap

- Bootstrap is a widely applicable and powerful statistical tool to quantify the uncertainty of a given estimator or statistical learning method.
- Principle of the bootstrap:
  - we obtain distinct data sets by repeatedly sampling observations from the original data set, with replacement;

    (rather than repeatedly obtaining independent data sets from the population)
  - we use the replicated data to estimate the parameter of interest or perform calculation of interest.

Bootstrap

- Example of bootstrapped data sets.

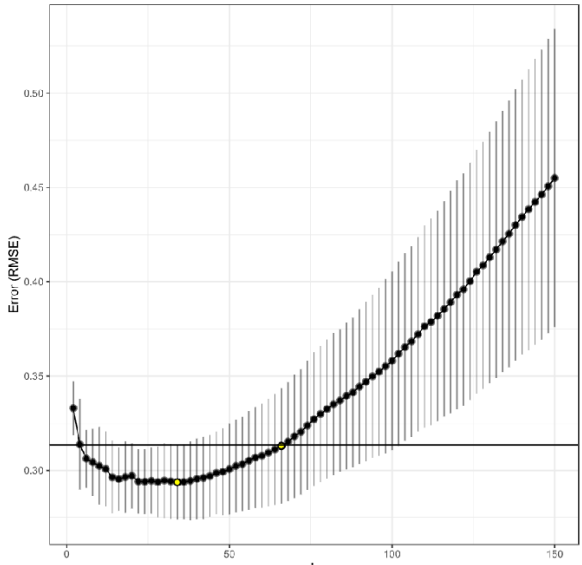


Choosing final tuning parameters

- We choose the final settings by:
  - quantifying model performance across sets of tuning parameters
  - pick the settings associated with the numerically best performance estimates.
- In general we prefer simpler models over more complex ones:
  - choosing tuning parameters based on the numerically optimal value may lead to models that are overly complicated
  - choose a simpler model that is within a certain tolerance of the numerically best value
  - 'one-standard error' rule:

    find the numerically optimal value and its corresponding s.e. and seek the model whose performance is within a single s.e. of the numerically best value.

Choosing final tuning parameters



Choosing between models

- Say the tuning parameters have been determined for each model.
- How do we choose between multiple models?
- This largely depends on the characteristics of the data and the type of questions being answered.

Choosing between models

- Scheme for finalizing the type of model:

  1. Start with several models that are the least interpretable and most flexible (e.g. boosted trees or support vector machines);

     Among many domains these models have a high likelihood of producing empirically optimal results (='gold standard').

  2. Investigate simpler models that are less opaque (e.g. not complete black boxes).

  3. Consider using the simplest model that reasonably approximates the performance of the more complex models.

- Reflection: How would we do this in a P&C pricing context (cfr. pricing analytics case study)?

Summary of models and their characteristics

Table A.1: A summary of models and some of their characteristics

Model	Allows $n < p$	Pre-processing	Interpretable	Automatic feature selection	# Tuning parameters	Robust to predictor noise	Computation time
Linear regression†	×	CS, NZV, Corr	✓	×	0	×	✓
Partial least squares	✓	CS	✓	○	1	×	✓
Ridge regression	×	CS, NZV	✓	×	1	×	✓
Elastic net/lasso	×	CS, NZV	✓	✓	1–2	×	✓
Neural networks	✓	CS, NZV, Corr	×	×	2	×	×
Support vector machines	✓	CS	×	×	1–3	×	×
MARS/FDA	✓		○	✓	1–2	○	○
K-nearest neighbors	✓	CS, NZV	×	×	1	○	✓
Single trees	✓		○	✓	1	✓	✓
Model trees/rules‡	✓		○	✓	1–2	✓	✓
Bagged trees	✓		×	✓	0	✓	○
Random forest	✓		×	○	0–1	✓	×
Boosted trees	✓		×	✓	3	✓	×
Cubist†	✓		×	○	2	✓	×
Logistic regression*	×	CS, NZV, Corr	✓	×	0	×	✓
{LQRM}DA*	×	NZV	○	×	0–2	×	✓
Nearest shrunken centroids*	✓	NZV	○	✓	1	×	✓
Naïve Bayes*	✓	NZV	×	×	0–1	○	○
C5.0*	✓		○	✓	0–3	✓	×

†regression only *classification only

Symbols represent affirmative (✓), negative (×), and somewhere in between (○)