Universität Potsdam

Institut für Informatik

Lehrstuhl Maschinelles Lernen

Model Evaluation

Tobias Scheffer

Overview

- Risk, empirical risk
- Precision, recall
- ROC curves
- Evaluation protocols
- Model selection

Learning and Evaluation

- Learning problem
  - Input: data $S = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$
  - Output: model $f_\theta : X \rightarrow Y$
- When model is applied, it is used to make predictions for new instances $\mathbf{x}$.
- How well will f_θ perform at application time?
  - What does “well” even mean?
  - How can it be determined?

Model Evaluation

- Central assumption about data: drawn according to single (unknown) distribution $p(\mathbf{x}, y)$.

- **"IID assumption"**: Instances $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$ are drawn independently and from an identical distribution.
- Independent: $p\left((\mathbf{x}_{i+j}, y_{i+j}) | (\mathbf{x}_i, y_i)\right) = p\left((\mathbf{x}_{i+j}, y_{i+j})\right)$.
- Identical distribution: $\forall i: (\mathbf{x}_i, y_i) \sim p(\mathbf{x}, y)$

Model Evaluation

- **“IID assumption”**: Instances $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ are drawn independently and from an identical distribution.
- Independent: $p\left((\mathbf{x}_{i+j}, y_{i+j}) | (\mathbf{x}_i, y_i)\right) = p\left((\mathbf{x}_{i+j}, y_{i+j})\right)$.
  - Counter-example: predicting ground motion during earthquakes.
  - $\mathbf{x}_i$: magnitude and epicenter of earthquake, sensor location, ground properties; y_i: ground acceleration at sensor location.
  - Observations at different sensor locations during the same earthquake are highly dependent.
  - Number of sensor stations often large, number of earthquakes is small; especially of high-magnitude earthquakes.
  - “Effective sample size” is much smaller than apparent sample size.

Model Evaluation

- **“IID assumption”**: Instances $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ are drawn independently and from an identical distribution.
- Independent: $p\left((\mathbf{x}_{i+j}, y_{i+j}) | (\mathbf{x}_i, y_i)\right) = p\left((\mathbf{x}_{i+j}, y_{i+j})\right)$.
  - Counter example: people who are surveyed at a random but fixed geographical location.
  - Consequence: a dependent sample contains less variance than an independent sample.

Model Evaluation

- **"IID assumption"**: Instances $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ are drawn independently and from an identical distribution.
- Independent: $p\left((\mathbf{x}_{i+j}, y_{i+j}) | (\mathbf{x}_i, y_i)\right) = p\left((\mathbf{x}_{i+j}, y_{i+j})\right)$.
  - Counter example: people who are surveyed at a random but fixed geographical location.
  - Consequence: a dependent sample contains less variance than an independent sample.
- Identical distribution: $\forall i: (\mathbf{x}_i, y_i) \sim p(\mathbf{x}, y)$
  - Counter example: first half of the data generated under laboratory conditions, second half collected "in the wild".
  - Consequence: model trained on laboratory data may perform poorly "in the wild".

Loss Function

- Loss function: How bad is it if the model predicts value $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ when the true value of the target variable is y_i?

$$\ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$$

- Example loss functions:
  - Zero-one loss (classification):

$$\ell_{0/1}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) = \begin{cases} 0 & \text{if } f_{\boldsymbol{\theta}}(\mathbf{x}_i) = y_i \\ 1 & otherwise \end{cases}$$

  - Quadratic loss (regression):

$$\ell_2(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) = (f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i)^2$$

  - Perceptron loss, hinge loss, ε-insensitive loss, …

Risk

- Risk of model f_θ: expected loss over underlying distribution $p(\mathbf{x}, y)$.
- Finite set Y (classification):

$$R(\theta) = E_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\ell(\mathbf{x}, y)] = \sum_{y\in Y} \int \ell(f_\theta(\mathbf{x}), y)p(\mathbf{x}, y)d\mathbf{x}$$

- Infinite Y (regression):

$$R(\theta) = E_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\ell(\mathbf{x}, y)] = \int \int \ell(f_\theta(\mathbf{x}), y)p(\mathbf{x}, y)d\mathbf{x}dy$$

- Expected zero-one loss (risk for zero-one loss function) is called **error rate.**
- 1-error rate is called **accuracy**.

Risk

- Risk of model f_θ: expected loss over underlying distribution $p(\mathbf{x}, y)$.
- Finite set Y (classification):

$$R(\theta) = E_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\ell(\mathbf{x}, y)] = \sum_{y \in Y} \int \ell(f_\theta(\mathbf{x}), y) p(\mathbf{x}, y) d\mathbf{x}$$

- Infinite Y (regression):

$$R(\theta) = E_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\ell(\mathbf{x}, y)] = \int \int \ell(f_\theta(\mathbf{x}), y) p(\mathbf{x}, y) d\mathbf{x} dy$$

- It is generally impossible to determine the risk:
  - $p(\mathbf{x}, y)$ is not known.
  - Generally impossible to integrate over all instances $\mathbf{x}$.

Empirical Risk

- Impossible to calculate risk

$$R(\theta) = E_{(\mathbf{x},y)\sim p(\mathbf{x},y)}\left[\ell(f_\theta(\mathbf{x}), y)\right]$$

- → Empirical risk: estimate on sample $S \sim p(\mathbf{x}, y)^n$.

$$\hat{R}_S(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta\,(\mathbf{x}_i, y_i))$$

- Empirical risk is a random variable; depends on the instances S that are drawn.
- If S is drawn **IID**, then it is governed by $p\big((\mathbf{x}_1, y_1) \times \cdots \times$

Estimators

- In statistics, an **estimator** is any rule for calculating an estimate of a quantity.
- A procedure for that determines the empirical risk is an estimator of the risk.

- An estimator is called unbiased if the expected value of the estimate is the true quantity:

$$\hat{R}(\theta) \text{is unbiased} \Leftrightarrow E_{S\sim p(\mathbf{x},y)^n}\left[\hat{R}_S(\theta)\right] = R(\theta)$$

- An estimator that is not unbiased has a bias:

$$B\left(\hat{R}(\theta)\right) = E_{S\sim p(\mathbf{x},y)^n}\left[\hat{R}_S(\theta)\right] - R(\theta)$$

Bias of the Empirical Risk

- Bias of the empirical risk:

$$B\left(\hat{R}(\theta)\right) = E_{S\sim p(\mathbf{x},y)^n}\left[\hat{R}_S(\theta)\right] - R(\theta)$$

- Empirical risk is unbiased estimator if:

$$E_{S\sim p(\mathbf{x},y)^n}\left[\hat{R}_S(\theta)\right] = R(\theta)$$

- Empirical risk is optimistic estimator if:

$$E_{S\sim p(\mathbf{x},y)^n}\left[\hat{R}_S(\theta)\right] - R(\theta) < 0$$

- Empirical risk is pessimistic estimator if:

$$E_{S\sim p(\mathbf{x},y)^n}\left[\hat{R}_S(\theta)\right] - R(\theta) > 0$$

Bias of the Empirical Risk

- Bias of the empirical risk:

$$B\left(\hat{R}(\theta)\right) = E_{S\sim p(\mathbf{x},y)^n}\left[\hat{R}_S(\theta)\right] - R(\theta)$$

- The bias is a systematical offset between risk and empirical risk.
- It can be caused by a particular experimental setting used to determine the empirical risk.
- Large bias: risk is systematically estimated too low or too high.

Variance of an Estimator

- Estimator $\hat{R}_S(\theta)$ has a variance

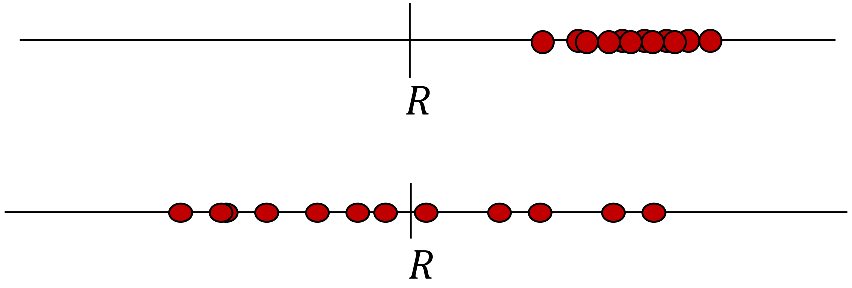
$$Var\left[\hat{R}_S(\theta)\right] = \mathrm{E}_{S\sim p(\mathbf{x},y)}\left[\left(E[R_S(\theta)] - \hat{R}_S(\theta)\right)^2\right]$$

- The variance is caused by the fact that the empirical risk is calculated on a finite sample.
- Zero-one loss: empirical risk $\hat{R}_S(\theta)$ follows binomial distribution with mean value $R(\theta)$.
- High variance: empirical risk is a crude estimate of the risk.
- The larger a sample the empirical risk is based on, the lower its variance becomes.

Bias and Variance of Empirical Risk

- Empirical risk $\hat{R}_S(\theta)$ determined repeatedly on multiple samples $S_1, \dots, S_k$

● Value of $\hat{R}_{S_i}$ for sample S_i



Large bias, small variance

Large variance,
small or no bias

Estimation Error

- Estimation error: expected quadratic difference between empirical risk and risk.

$$\mathrm{E}_{S\sim p(\mathbf{x},y)^n}\left[\left(\hat{R}_S(\theta)-R(\theta)\right)^2\right]$$

- Can be decomposed into bias and variance

$$\begin{aligned}
&\mathrm{E}_{S\sim p(\mathbf{x},y)^n}\left[\left(\hat{R}_S(\theta)-R(\theta)\right)^2\right] \\
&= \mathrm{E}\left[\hat{R}_S(\theta)^2-2R(\theta)\hat{R}_S(\theta)+R(\theta)^2\right] \\
&= \mathrm{E}\left[\hat{R}_S(\theta)^2\right]-2R(\theta)\mathrm{E}\left[\hat{R}_S(\theta)\right]+R(\theta)^2 \\
&= \mathrm{E}\left[\hat{R}_S(\theta)\right]^2-2R(\theta)\mathrm{E}\left[\hat{R}_S(\theta)\right]+R(\theta)^2+\mathrm{E}\left[\hat{R}_S(\theta)^2\right]-\mathrm{E}\left[\hat{R}_S(\theta)\right]^2 \\
&= \left(\mathrm{E}\left[\hat{R}_S(\theta)\right]-R(\theta)\right)^2+\mathrm{E}\left[\hat{R}_S(\theta)^2\right]-\mathrm{E}\left[\hat{R}_S(\theta)\right]^2 \\
&= Bias\left[\hat{R}(\theta)\right]^2+Var\left[\hat{R}(\theta)\right]
\end{aligned}$$

Algebraic formula for the variance

Overview

- Risk, empirical risk
- Precision, recall
- ROC curves
- Evaluation protocols
- Model selection

Alternative Measures to Risk

- Risk is not always a meaningful measure.
- Not always possible to specify a meaningful loss function
  - Mine detector: what is the cost of exploding?
  - On the other hand, a mine detector that always says “there could be a mine here” is useless.
- Error rate / accuracy are not meaningful for rare classes.
  - Earth quake prediction tool that always says “there will be no earthquake today” has accuracy of >99.9% (in most countries).

Alternative Measures to Risk

- Alternative performance measures for binary classification.
- Let decision function $f_\theta(\mathbf{x})$ return continuous value.
- Decision rule for binary classification: $y_\theta(\mathbf{x}) =$
$$\begin{cases} +1 & \text{if } f_\theta(\mathbf{x}) \geq \theta_0 \\ -1 & \text{if } f_\theta(\mathbf{x}) < \theta_0 \end{cases}$$
- By adjusting threshold θ_0 decision rule can be made more sensitive or more conservative.
- We will now study measures that quantify how well the decision function separates positive from negative instances, independent of any threshold value θ_0.
    - Precision-recall curves
    - ROC curves

Overview

- Risk, empirical risk
- Evaluation protocols
- Model selection
- Precision, recall
- ROC curves

Precision and Recall

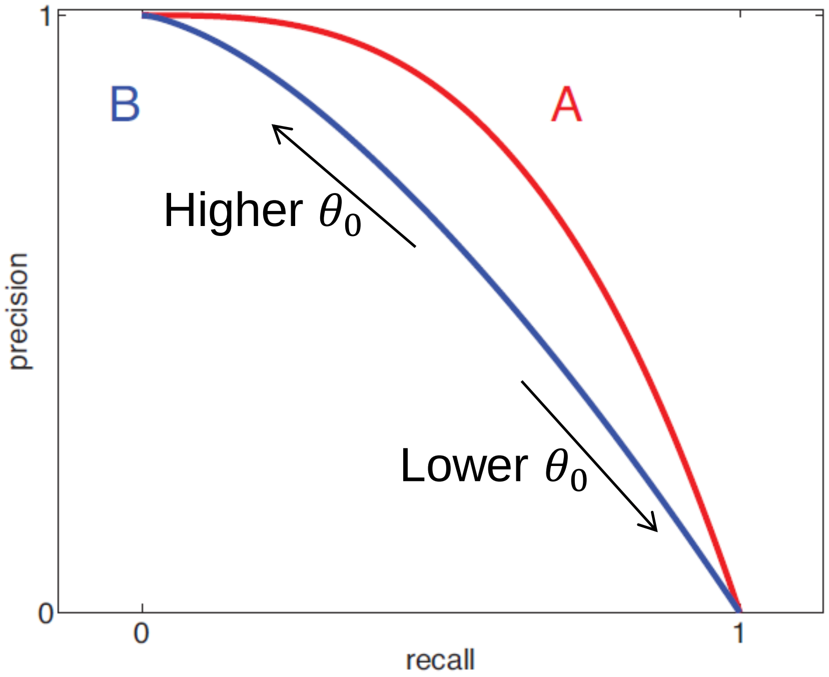
- Alternative performance measure for binary classification.
  - Example: medical diagnostics system for rare disease.
  - Patient $\mathbf{x}_i$ has disease if $y_i = +1$.
  - Classifier diagnoses disease for patient $\mathbf{x}$ if $y_\theta(\mathbf{x}_i) = +1$.
- True positives:
  - Patient has disease ($y_i = +1$), classifier recognizes ($y_\theta(\mathbf{x}_i) = +1$)
- False positives:
  - Patient is healthy ($y_i = -1$), but classifier diagnoses disease ($y_\theta(\mathbf{x}_i) = +1$)
- True negatives:
  - Patient is healthy ($y_i = -1$), classifier recognizes ($y_\theta(\mathbf{x}_i) = -1$)
- False negatives:
  - Patient has disease ($y_i = +1$), classifier misses ($y_\theta(\mathbf{x}_i) = -1$)

Precision and Recall

- Let n_{TP} be the number of true positives.
- Let n_{FP} be the number of false positives.
- Let n_{TN} be the number of true negatives.
- Let n_{FN} be the number of false negatives.
- Precision: $P = \frac{n_{TP}}{n_{TP}+n_{FP}}$
  - Rate of true positives among all instances that are classified as positives
  - Answers: “How accurate is classifier when it says +1?”
- Recall: $R = \frac{n_{TP}}{n_{TP}+n_{FN}}$
  - Rate of true positives among all positive instances
  - Answers: “How many of the positive instances does the classifier detect?”

Precision-Recall Curves

- Evaluates decision function $f_\theta(\mathbf{x})$ independent of threshold θ_0.
- Shows which pairs of precision and recall can be obtained by varying threshold θ_0.
- Each point on the curve is a classification rule with a particular values of θ_0.
- Which decision function is better – A or B?



Example: Malware Detection

- Instances are client computers
- Error rate $E_{(\mathbf{x},y)}\left[\ell_{0/1}(\mathbf{x},y)\right]$:
  - Rate at which computers are misclassified as infected / not infected.
- If 0.1% of all computers are infected:
  - An error rate of 0.1% may mean that no infection is detected.
  - An error rate of 0.2% may mean all alarms are false alarms.

Example: Malware Detection

- Precision: $P = \frac{n_{TP}}{n_{TP}+n_{FP}}$
  - Rate of actual infections among all alarms.
  - $P = 80\%$ means: 20% of all alarms are false alarms.
  - Low precision means the tool is annoying.
- Recall: $R = \frac{n_{TP}}{n_{TP}+n_{FN}}$
  - Proportion of malware that is detected.
  - $R = 60\%$ means: 40% of all malware goes unnoticed.
  - Low recall means the tool is useless.
- Precision and recall are good performance measures for malware detection.
- Caveat: precision and recall of a fixed model change when the class ratio changes.

F Measures

- F_α measures combine precision and recall values into single value:

$$F_\alpha = \frac{n_{TP}}{\alpha(n_{TP} + n_{FP}) + (1 - \alpha)(n_{TP} + n_{FN})}$$

- $\alpha = 1$: Precision
- $\alpha = 0$: Recall
- $\alpha = 0.5$: “F-measure”, harmonic mean of precision and recall.
- Alternative definition: F_β measures.
  - Relationship: $\alpha = \frac{1}{1+\beta}$

F Measures

- F_α measures are not the aggregate of the precision-recall curve but of a single point on that curve.
- F_α measures are highly dependent on that point, determined by the decision threshold θ_0.
- There is rarely (never?) a good motivation for maximizing F_α measures in machine learning.
- However, if maximizing F_α is the goal, it is crucial to adjust θ_0 for maximal F_α.
- The value of θ_0 that maximizes F_α depends on the class ratio.
- Measuring F_α for a default threshold such as $\theta_0 = 0$ makes no sense at all.

Overview

- Risk, empirical risk
- Evaluation protocols
- Model selection
- Precision, recall
- ROC curves

ROC Analysis

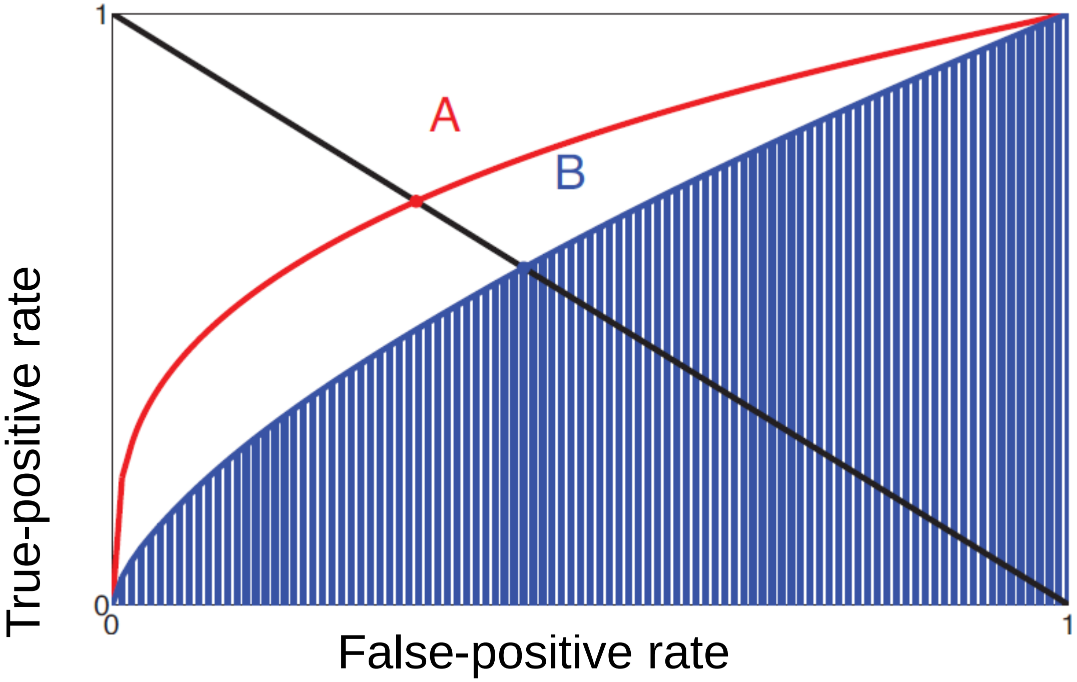
- Alternative performance measure for binary classification.
  - Example: diagnosis of rare disease.
  - Patient $\mathbf{x}_i$ has disease if $y_i = +1$.
  - Classifier diagnoses disease for patient $\mathbf{x}$ if $y_\theta(\mathbf{x}_i) = +1$.
- True-positives:
  - Patient has the disease ($y_i = +1$), classifier recognizes it ($y_\theta(\mathbf{x}_i) = +1$)
- False positives:
  - Patient is healthy ($y_i = -1$), but classifier diagnoses disease ($y_\theta(\mathbf{x}_i) = +1$)
- True negatives:
  - Patient is healthy ($y_i = -1$), classifier recognizes ($y_\theta(\mathbf{x}_i) = -1$)
- False negatives:
  - Patient has disease ($y_i = +1$), classifier misses ($y_\theta(\mathbf{x}_i) = -1$)

ROC Analysis

- Let n_{TP} be the number of true positives.
- Let n_{FP} be the number of false positives.
- Let n_{TN} be the number of true negatives.
- Let n_{FN} be the number of false negatives.
- True-positive rate (recall): $r_{TP} = \frac{n_{TP}}{n_{TP}+n_{FN}}$
  - Rate of true positives among all positive instances
  - Answers: “How many of the positive instances does the classifier detect?”
- False-positive rate: $r_{FP} = \frac{n_{FP}}{n_{FP}+n_{TN}}$
  - Rate of false positives among all instances that are really negatives.
  - Answers: “How many of the negative instances does the classifier misclassify as positive?”

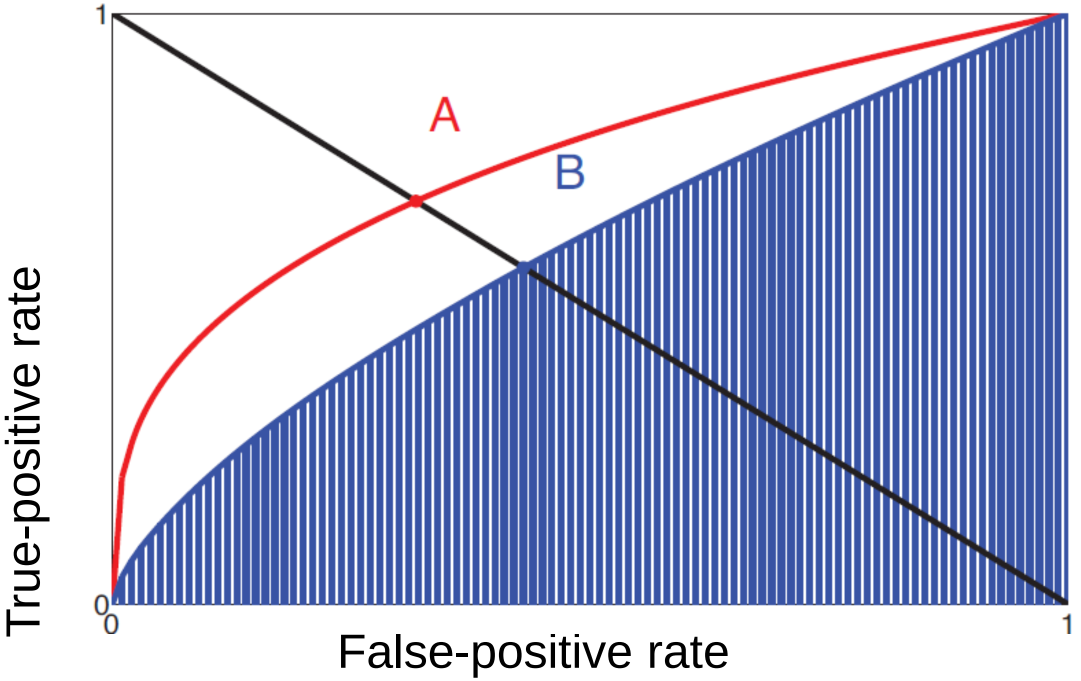
ROC Analysis

- Alternative measure of how well the decision function separates positive from negative instances, independent of any threshold value θ_0.



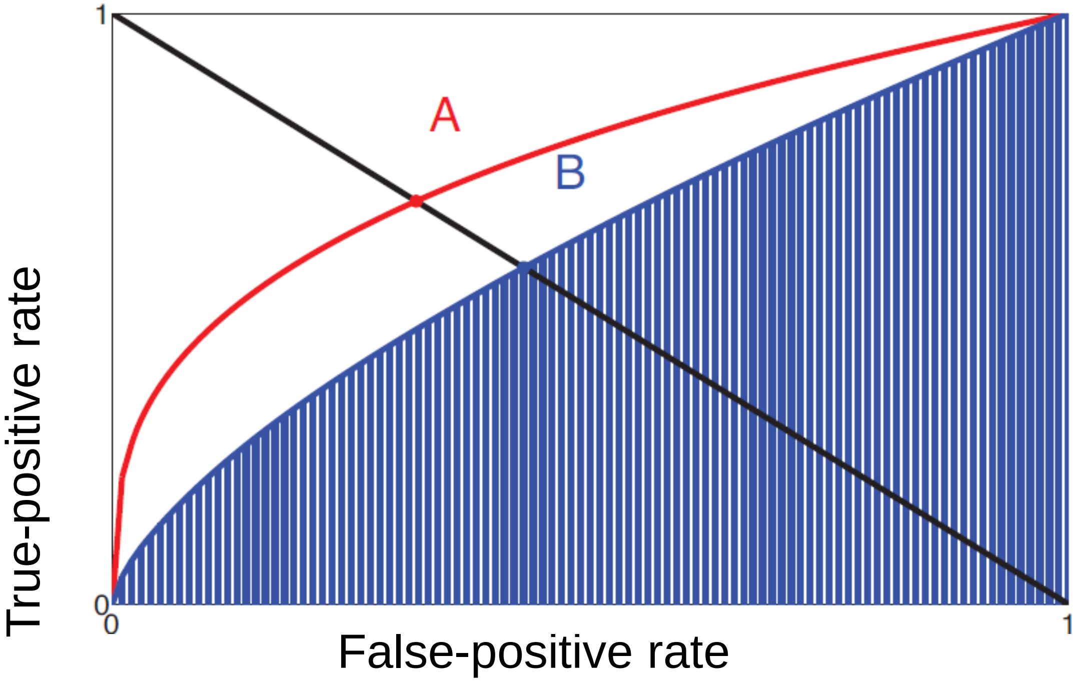
ROC Analysis

- Each curve characterizes a decision function f_θ.
- Each point is a classification rule for a value of θ_0.
- Which is better, A or B?
- $r_{TP} = \frac{n_{TP}}{n_{TP}+n_{FN}}$
- $r_{FP} = \frac{n_{FP}}{n_{FP}+n_{TN}}$



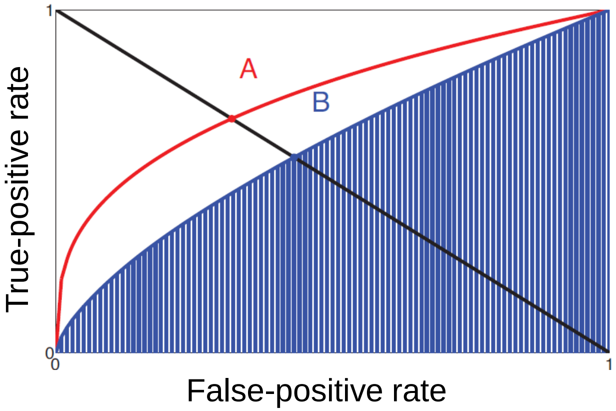
ROC Analysis

- Equal error rate (EER): value $r_{TP} = 1 - r_{FP}$.
- Scalar aggregate of curve: Area under ROC curve (AUC).



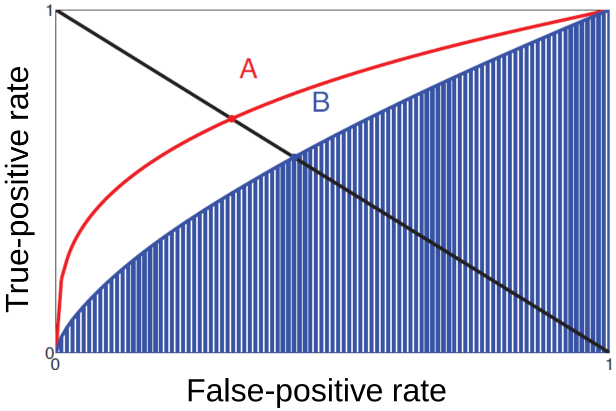
ROC Analysis

- Area under the ROC curve (AUC):
  - Let $\mathbf{x}_+$ be a randomly drawn positive instance.
  - Let $\mathbf{x}_-$ be a randomly drawn negative instance.
  - $AUC(\theta) = P(f_\theta(\mathbf{x}_+) > f_\theta(\mathbf{x}_-))$.
- AUC = Probability that randomly drawn positive instance has higher score than random negative instance.



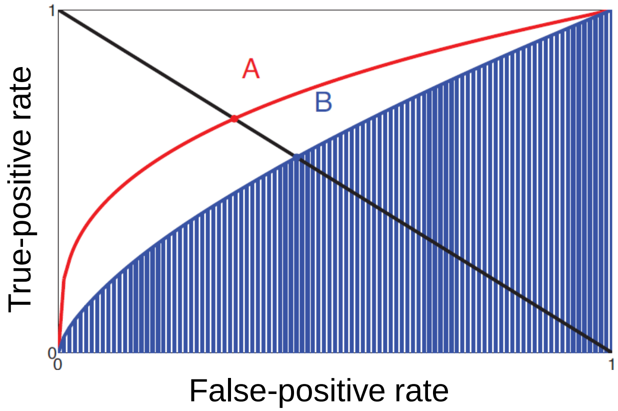
ROC Analysis

- ROC curves are invariant with respect to class ratio.
- When the ratio of positive to negative instances changes over time (and model stays fixed):
  - The error rate generally changes;
  - Precision and recall generally change;
  - The ROC curve (and AUC) remain the same.
- ROC analysis is great when the class ratio can change.
  - E.g., medical diagnostics during epidemic.



ROC Analysis

- ROC analysis is often used
  - When positive instances are rare (accuracy of 99.9% is meaningless if positive class is extremely rare)
  - When class ratio is unknown (probability of stepping on a mine varies by country).
  - When class ratio changes over time (probability that a patient has influenza varies seasonally).



Example: Medical Diagnosis

- True-positive rate (recall): $r_{TP} = \frac{n_{TP}}{n_{TP}+n_{FN}}$
  - Proportion of infected persons that are detected.
  - $r_{TP} = 60\%$ means: 40% of all infactions go unnoticed.
  - Low recall means the diagnostic tool is useless.
- False-positive rate: $r_{FP} = \frac{n_{FP}}{n_{FP}+n_{TN}}$
  - Rate at which healthy individuals are diagnosed as infected.
  - $r_{FP} = 5\%$ means that 5% of all healthy patients have to undergo further diagnostic procedures or potentially unneeded treatment.
  - High false-positive rate means that many patients will be sent to further diagnostic procedures.
- True-positive and false-positive rate are invariant with respect to class ratio.

Overview

- Risk, empirical risk
- Precision, recall
- ROC curves
- Evaluation protocols
- Model selection

Evaluation Protocols

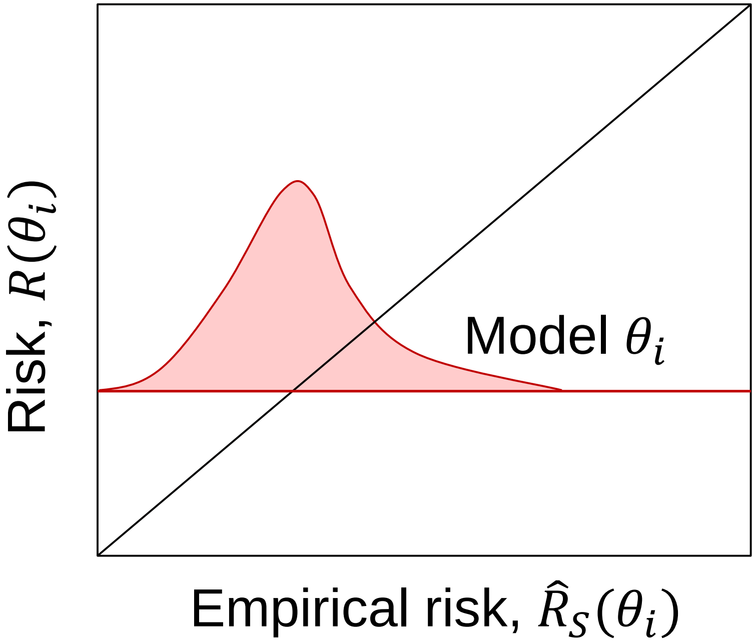
- Usually, model f_θ is not given and evaluation data cannot be drawn from $p(\mathbf{x}, y)$.
- Typical case, data $S = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ and learning method are given.
- Data S have to be used for training and evaluation.
- Desired output: model f_θ and risk estimate.

Evaluation Protocols

- Can we first train model f_θ on S and then evaluate the model on the same data?
- Will $\hat{R}_S(\theta)$ be unbiased, optimistic, or pessimistic?

Evaluation Protocols

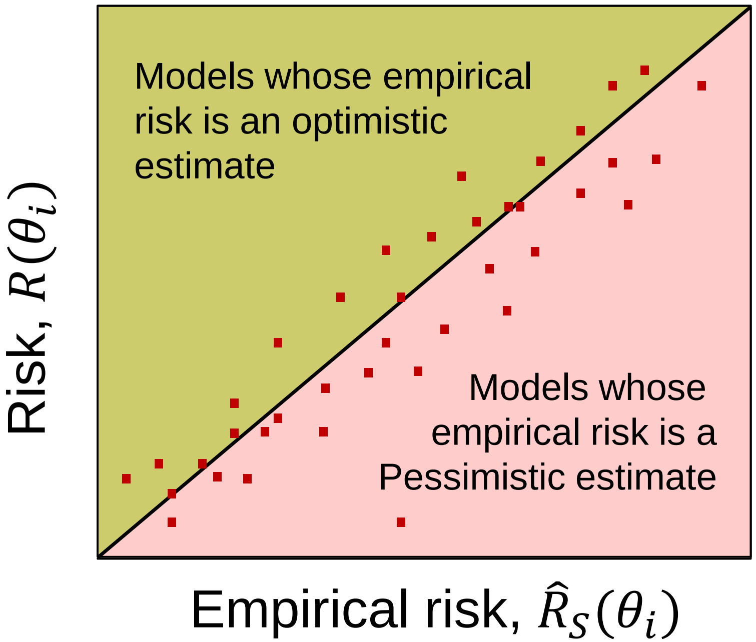
- Every model $\theta_i \in \Theta$ has a risk $R(\theta_i)$.
- Its empirical risk $\hat{R}_S(\theta_i)$ follows a distribution with mean value $R(\theta_i)$.



Evaluation Protocols

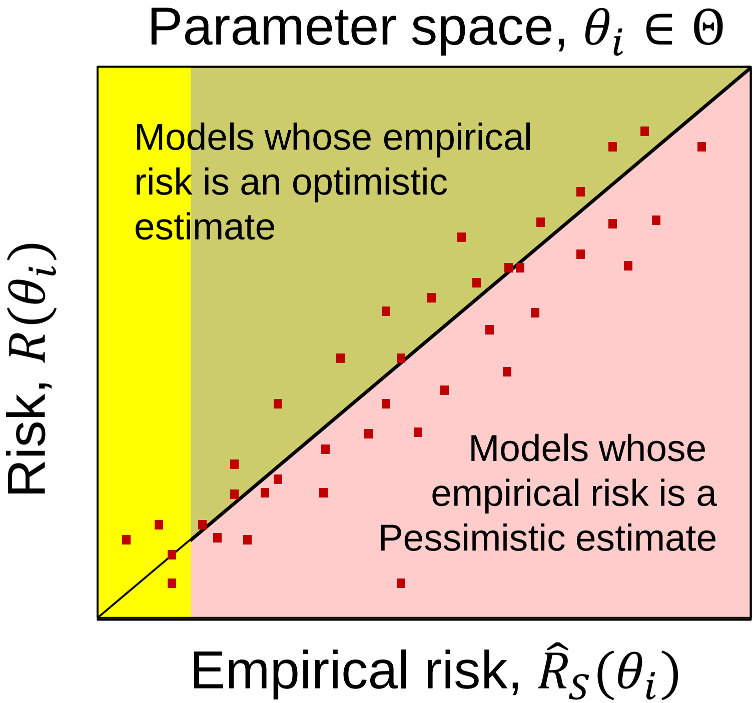
- Some models get lucky (upper-left area), some are unlucky (lower-right area).

Parameter space, $\theta_i \in \Theta$



Evaluation Protocols

- Learning algorithm will choose a model with small empirical risk (on the far left).
- In this area, most models‘ empirical risk is an optimistic estimate.

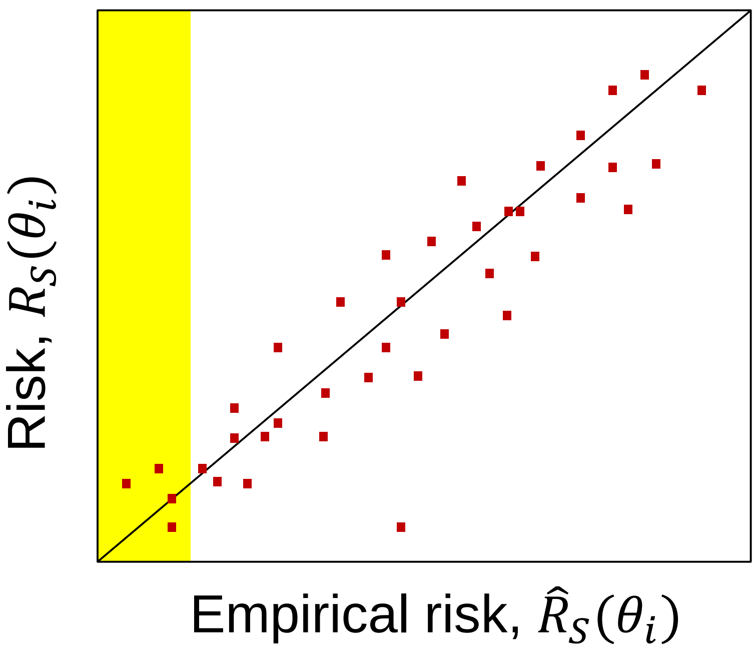


Parameter space, $\theta_i \in \Theta$
Models whose empirical risk is an optimistic estimate
Models whose empirical risk is a Pessimistic estimate
Risk, $R(\theta_i)$
Empirical risk, $\hat{R}_S(\theta_i)$

Evaluation Protocols

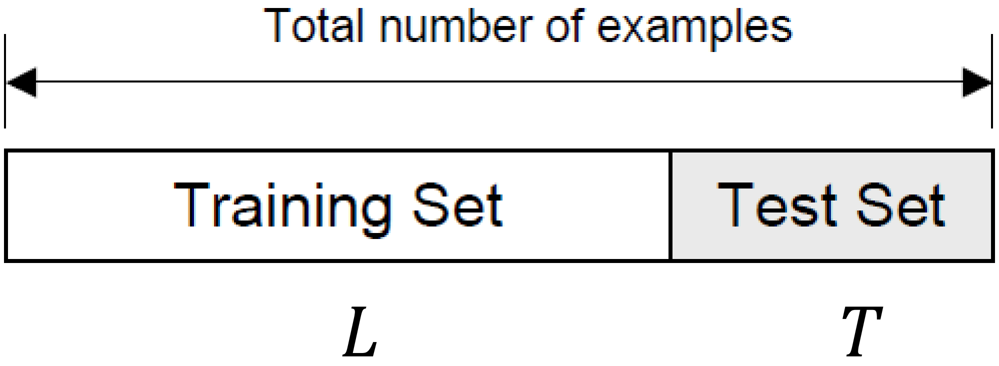
- Learning algorithm will choose a model with small empirical risk (on the far left).
- For those θ_* on the left: $E_S\left[\hat{R}_S(\theta_*)\right] < R(\theta_*)$ (otherwise they would be further right).
- This is called **selection bias**.
- **Empirical risk on training data is optimistic.**

Parameter space, $\theta_i \in \Theta$



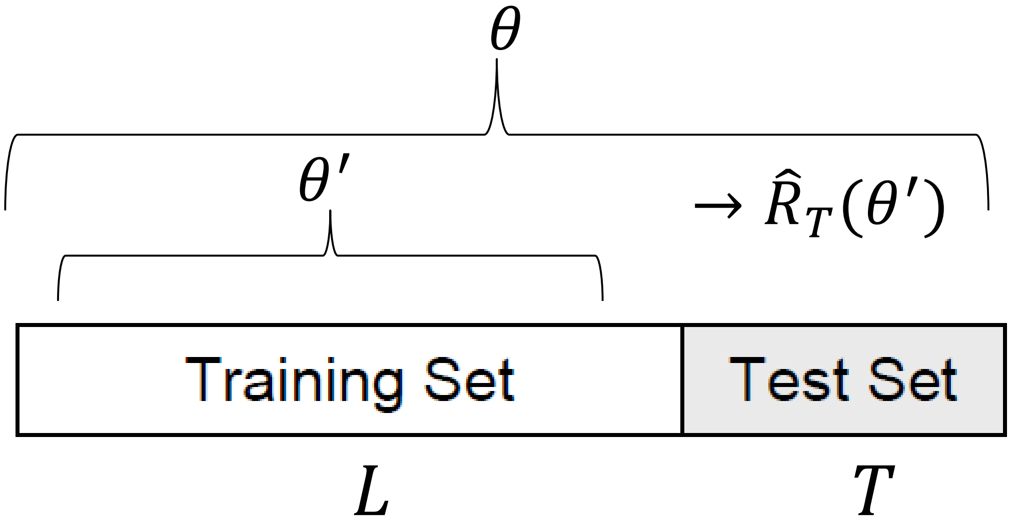
Holdout Testing

- Idea: error estimation on independent test data
- Given: data $S = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$
- Divide the data into
  - Training data $L = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$and
  - Test data $T = (\mathbf{x}_{m+1}, y_{m+1}), \dots, (\mathbf{x}_n, y_n)$



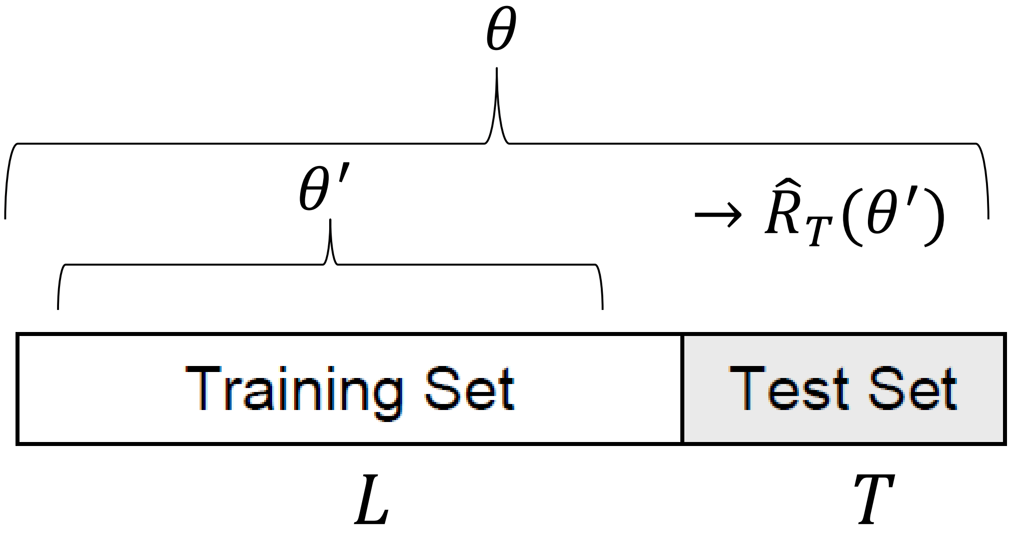
Holdout Testing

- Start learning algorithm with data L and obtain model f_{θ}, from it.
- Determine empirical risk $\hat{R}_T(\theta')$ from data T.
- Start learning algorithm with all data S and obtain Model f_θ from it.
- Output: model f_θ & $\hat{R}_T(\theta')$ as the estimator of $R(\theta)$.



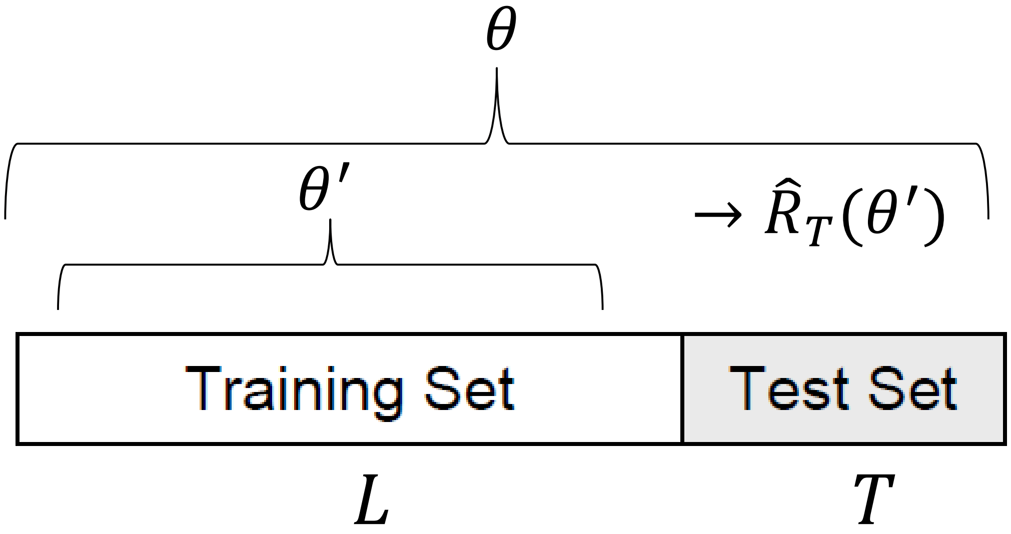
Holdout Testing: Analysis

- Is the estimator $\hat{R}_T(\theta')$ of the risk of model $R(\theta)$
  - unbiased,
  - optimistic,
  - pessimistic?
- Hint: the more training data, the better the model.



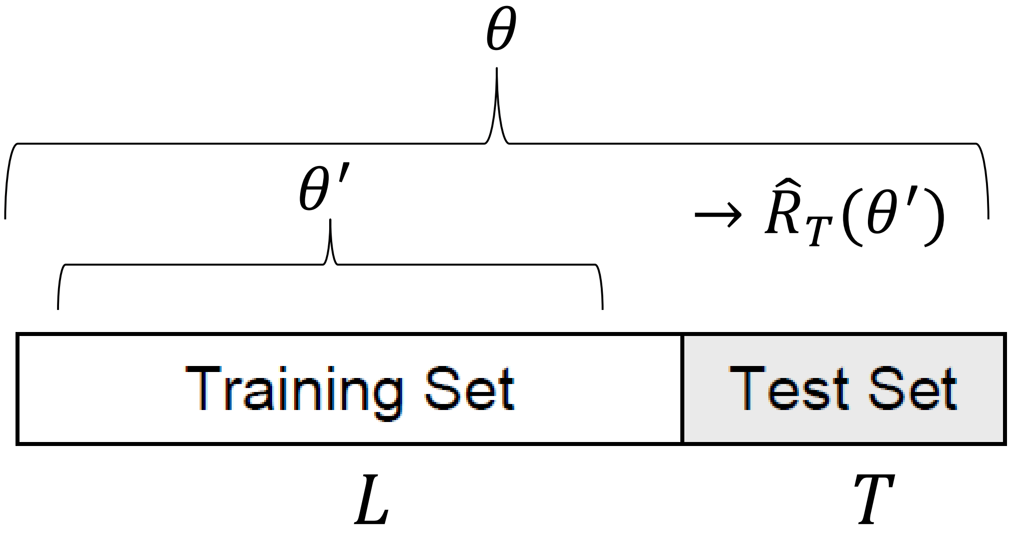
Holdout Testing: Analysis

- Estimate $\hat{R}_T(\theta')$ is obtained on a small part of the available data.
- Therefore, its variance is relatively high, especially if the overall sample is small.
- Holdout testing is used in practice for large available samples.



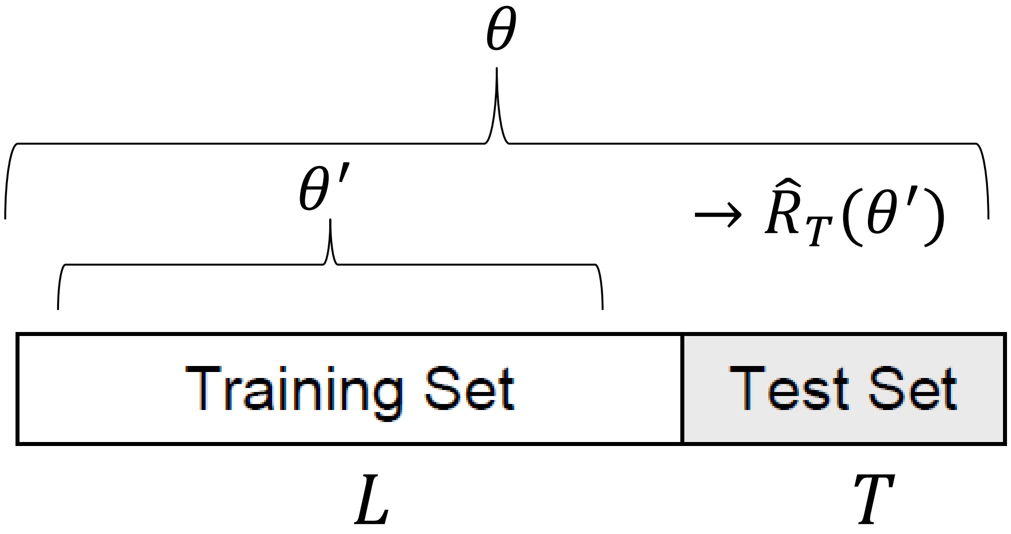
Holdout Testing: Analysis

- Using empirical risk $\hat{R}_T(\theta')$ is an **pessimistic** estimator of the risk $R(\theta)$.
- Because θ' is trained with fewer training instances than θ.



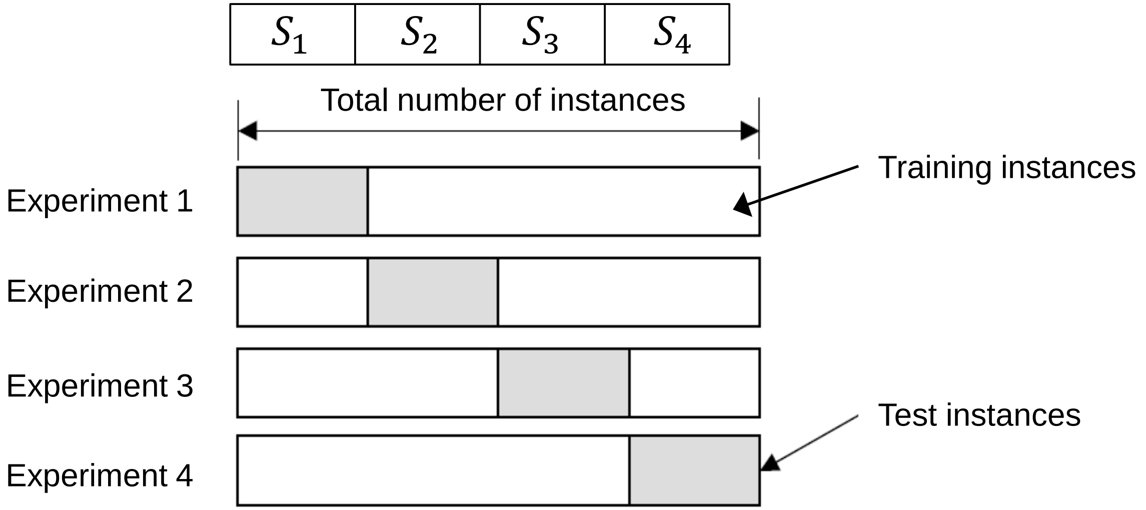
Holdout Testing: Analysis

- One could instead return model θ'.
- Empirical risk $\hat{R}_T(\theta')$ would be an unbiased estimate of $R(\theta')$.
- But since θ' was trained on fewer data, it would result in an inferior model.



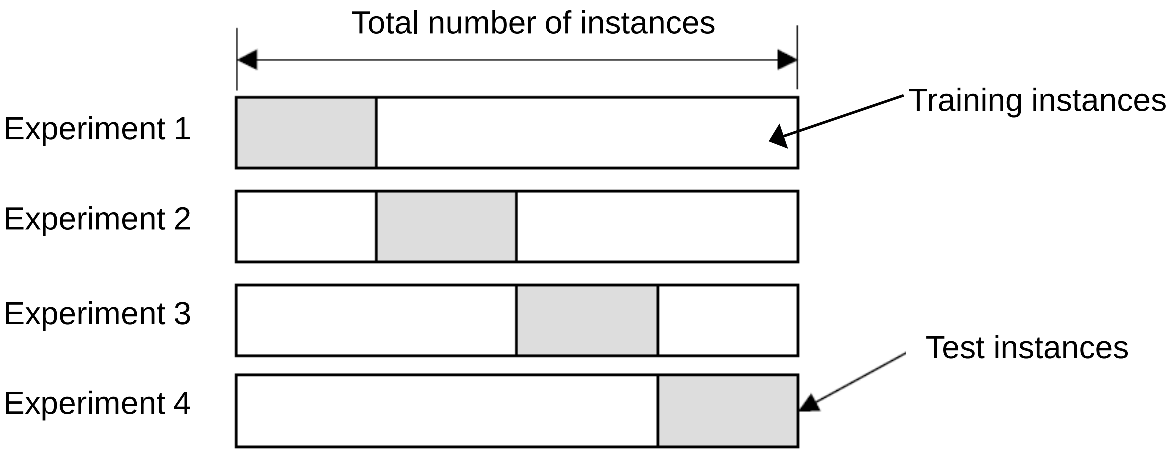
K-Fold Cross Validation

- Given: data $S = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$
- Partition S into k equally sized portions $S_1, \dots, S_k$.
- Repeat for $i = 1 \dots k$
  - Train f_{θ_i} with training set $S = S \backslash S_i$.
  - Calculate empirical risk $\hat{R}_{S_i}(\theta_i)$ on S_i.
- Calculate average $\hat{R}_S = \frac{1}{k} \sum_i \hat{R}_{S_i}(\theta_i)$



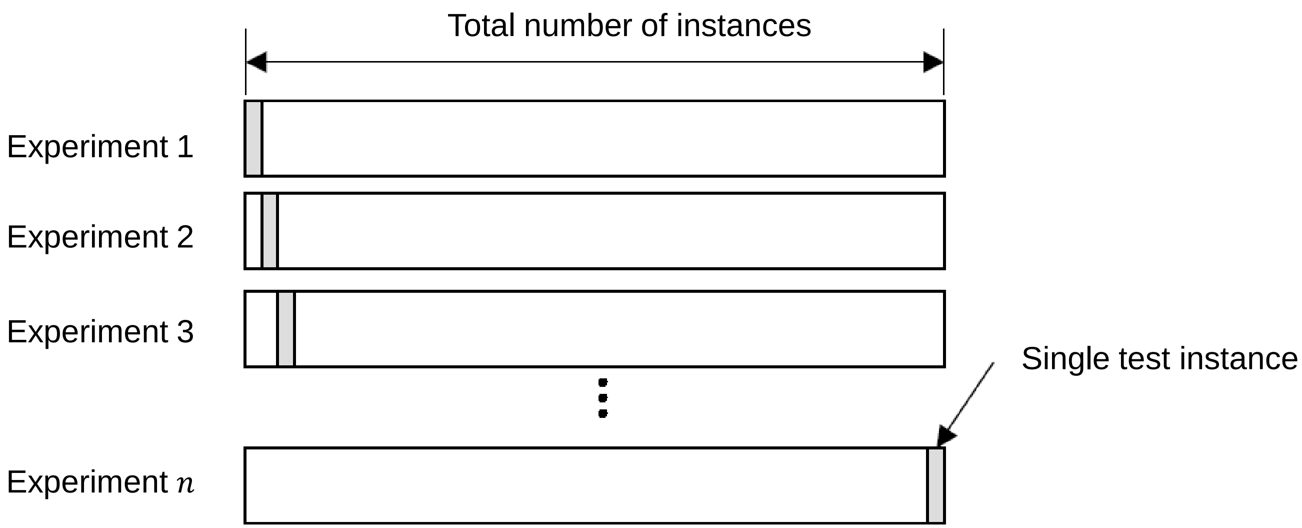
Cross Validation

- Then, train f_θ on all data S.
- Return model f_θ and estimator $\hat{R}_S$.



Leave-One-Out Cross Validation

- Special case $k = n$ is also called *leave-one-out* error estimation



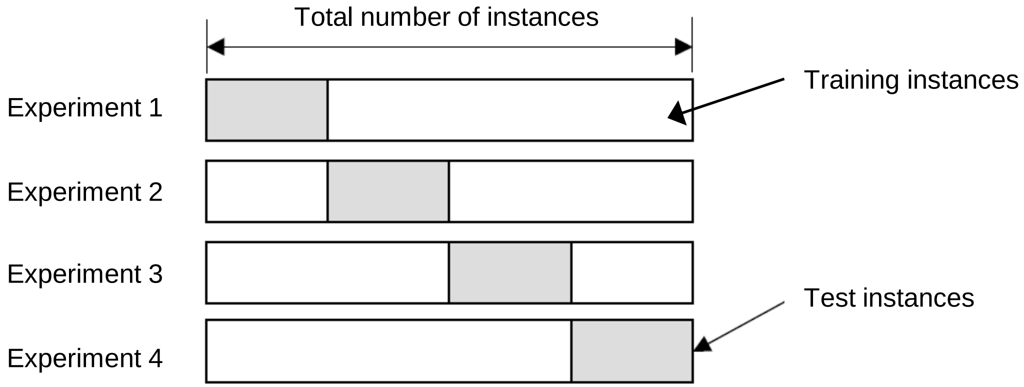
Cross Validation: Analysis

- Is the estimator
  - optimistic / pessimistic / unbiased?

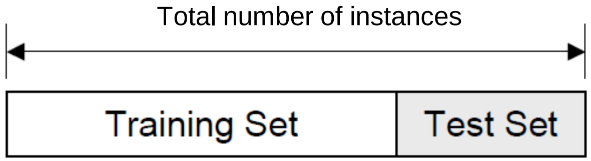
Cross Validation: Analysis

- Is the estimator
  - optimistic / pessimistic / unbiased?
- Estimator is slightly pessimistic:
  - Model f_{θ_i} is trained on a $(k-1)/k$-th fraction of the available data.
  - Model f_θ is trained on the entire data.

Cross Validation



Holdout

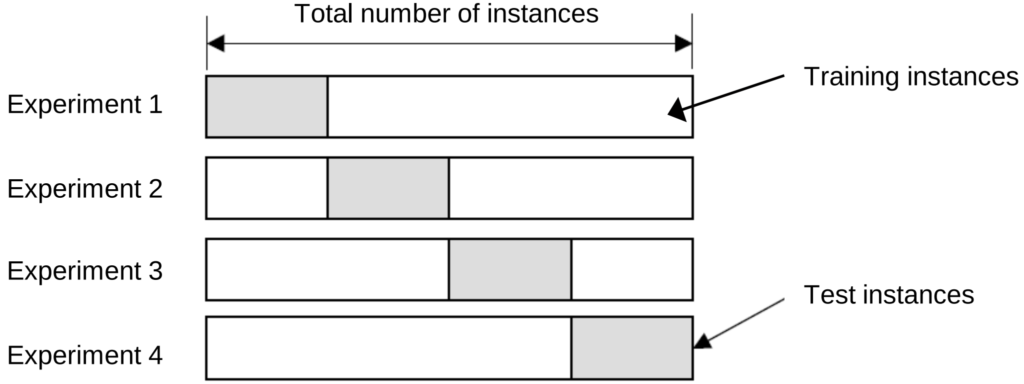


Intelligent Data Analysis

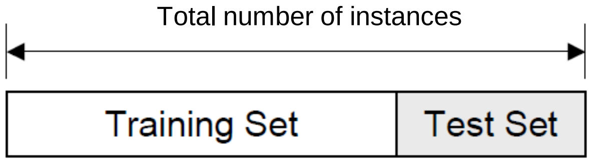
Cross Validation: Analysis

- Bias/Variance compared to holdout testing?
- Variance is lower than with holdout testing
  - Averaging over several holdout experiments reduces the estimator’s variance.
  - All data is incorporated into the estimator.
- Bias similar to holdout testing, depending on the split ratios.

Cross Validation



Holdout



Overview

- Risk, empirical risk
- Precision, recall
- ROC curves
- Evaluation protocols
- Model selection

Model Selection

- Compare several different learning approaches
  - Should one use decision trees?
  - SVMs? Logistic Regression?
- Set regularization parameter for a learning approach
  - For instance, set value for λ for regularized empirical risk minimization.

Model Selection: Example

- Regularization parameter λ in optimization criterion

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_i \ell(f_\theta(\mathbf{x}_i), y_i) + \lambda ||\theta||^2 \qquad \lambda = ?$$

- (Hyper)parameters that specify the model class; e.g. the degree for polynomial regression

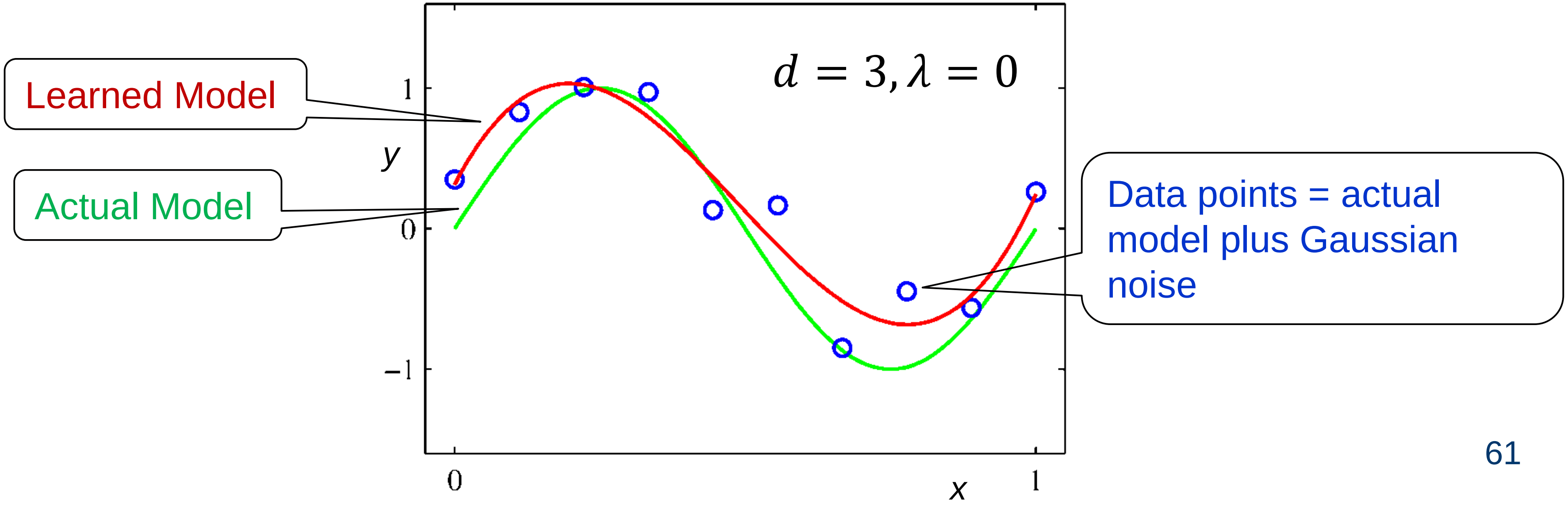
$$f_{\boldsymbol{\theta}}(x) = \sum_{j=0}^{d} \theta_j x^j \qquad d = ?$$

- Desired output: hyperparameter (λ, d), model f_θ, and estimate of the model's risk.
- How do we use available data to achieve this?

Example: Polynomial Regression

- Polynomial model of degree d: $f_\theta^d(x) = \sum_{j=0}^{d} \theta_j x^j$

- Regularized empirical risk minimization: $\theta^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(f_\theta^d(x_i) - y_i\right)^2 + \lambda\left|\left|\theta\right|\right|^2$



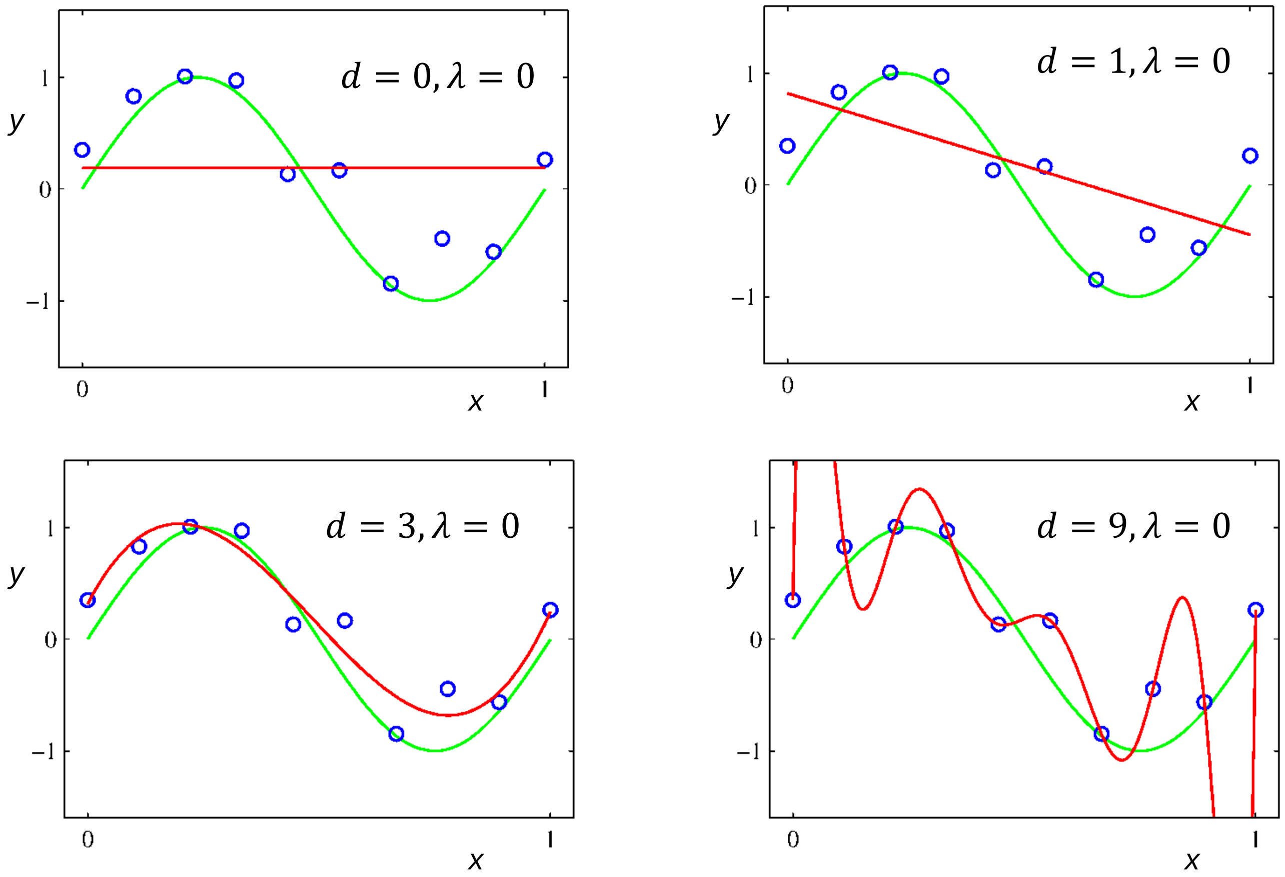
Learned Model

Actual Model

Data points = actual model plus Gaussian noise

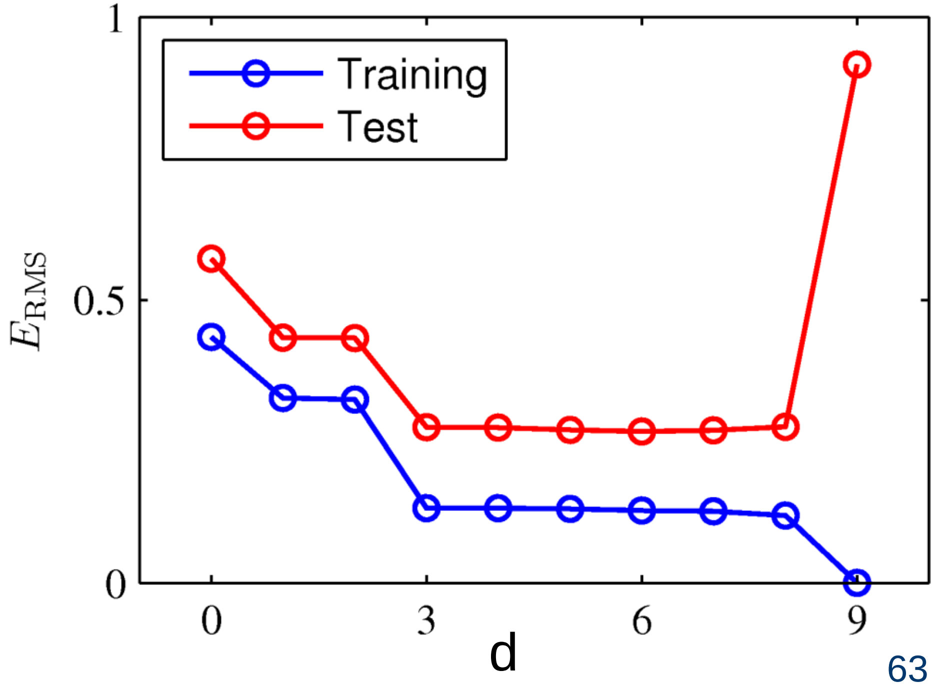
Polynomial Regression

- Success of the learning depends on the selected polynomial degree d, which controls the complexity of the model.



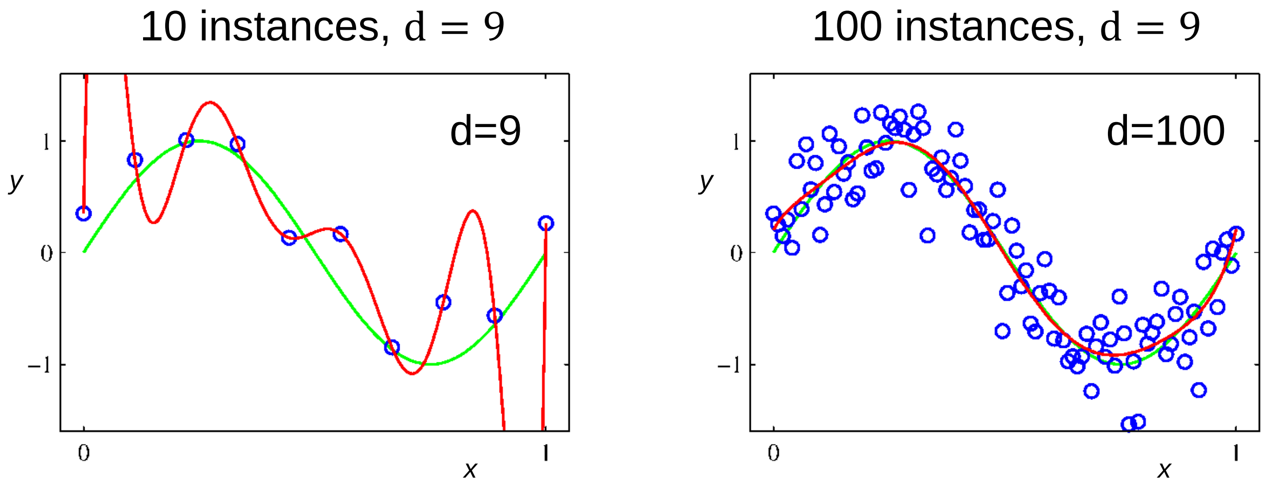
Polynomial Regression: Empirical Risk on Training vs. Test Sample

- Empirical risk on training vs. test data for different polynomial degrees.
- “Overfitting”: empirical risk on training data decreases as d is increased. Empirical risk on test data has a minimum, then increases again.



Example: Polynomial Regression

- If more data are available, more complex models can be fitted.



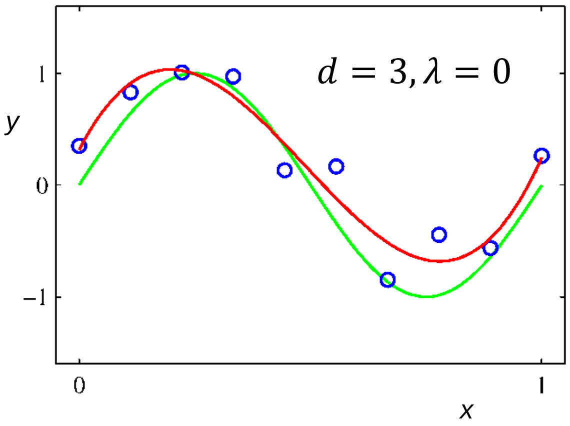
10 instances, d = 9

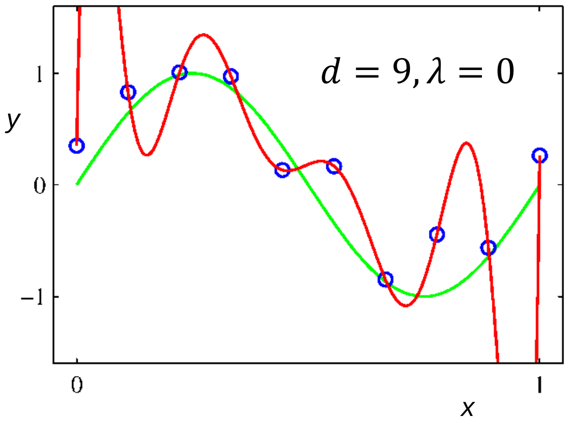
100 instances, d = 9

- Given fixed amount of data, optimal d has to be found.

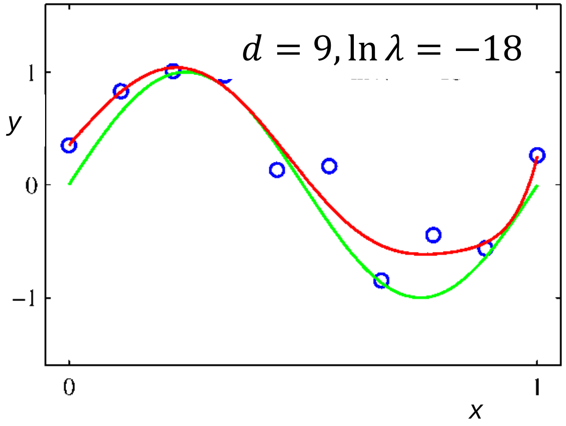
Example: Polynomial Regression

- Regularization factor λ has a similar effect to d.



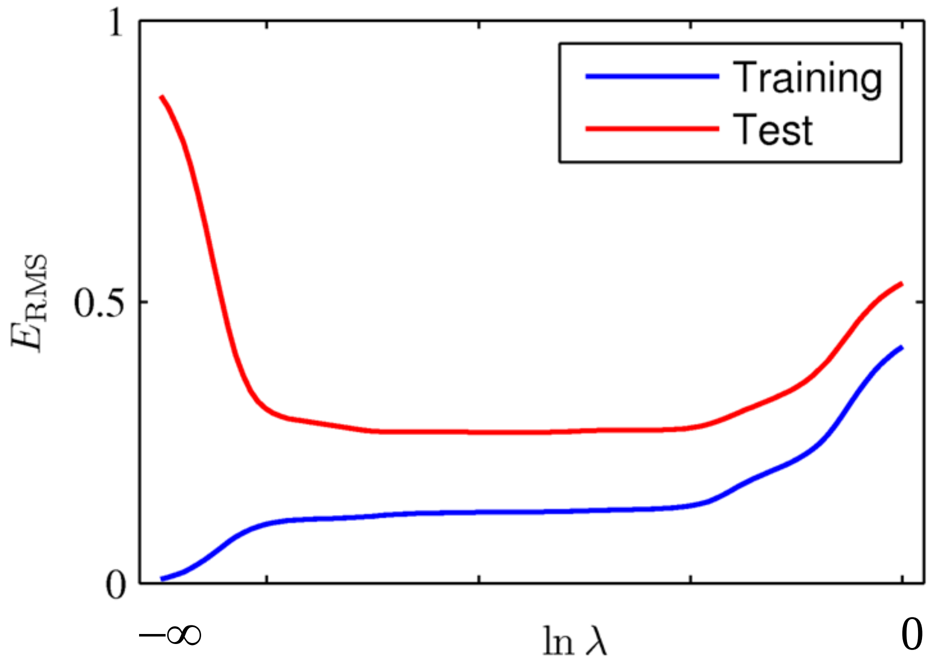


- Both λ and d constrain the model complexity.



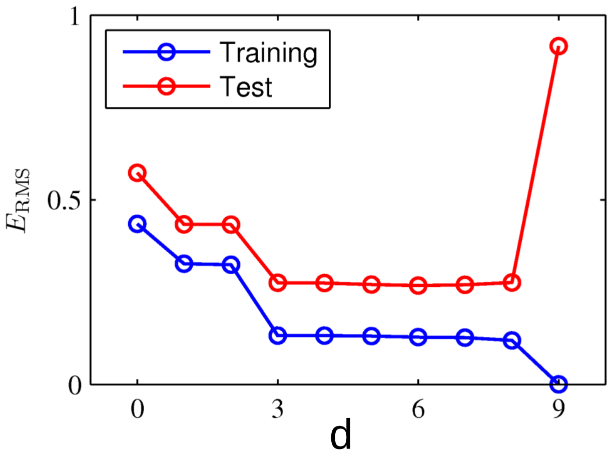
Regularized Polynomial Regression

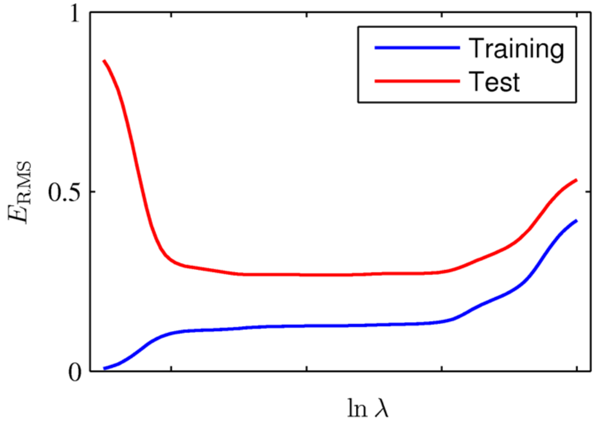
- Empirical risk on training vs. test sample.
- Empirical risk on training sample decreases when regularization decreases.
- There is a regularization factor that minimizes the risk.



Regularized Polynomial Regression

- Regularizer acts like a limitation on the model complexity and prevents overfitting.
- In practice it is best to control model complexity through regularization (direct parameters like the polynomial degree often are not available).
- Regularizer has to be tuned on available data.



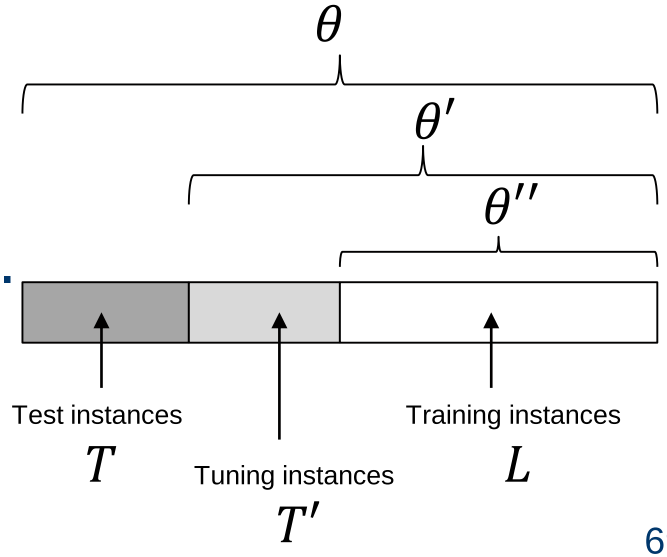


Model Selection, Setting Hyperparameters

- Desired output: hyperparameter (λ, d), model f_θ, and estimate of the model's risk.
- Idea: Iterate over values of (λ, d), train model, evaluate; take best values and train final model.
- Cannot tune hyperparameters on training data because low regularization leads to low empirical risk on training data but high risk on test data.
- Evaluating multiple models (for different values of λ, d) on the same test set results in an optimistic bias.
- Therefore, triple or nested cross validation.

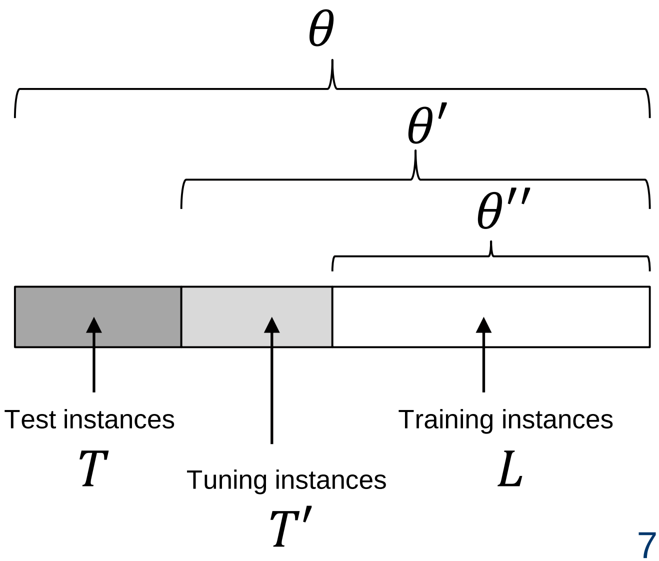
Triple Cross Validation

- Iterate over all values of the hyperparameters λ (grid search)
  - Train model $f_{\theta''}^{\lambda}$ on L.
  - Evaluate $f_{\theta''}^{\lambda}$on T' by calculating $\hat{R}_{T'}(f_{\theta''}^{\lambda})$
- Use hyperparameter λ^* that gave lowest $\hat{R}_{T'}(f_{\theta''}^{\lambda^*})$.
- Train model $f_{\theta'}^{\lambda^*}$ on $L \cup T'$.
- Determine $\hat{R}_T(f_{\theta'}^{\lambda^*})$.
- Train model $f_{\theta}^{\lambda^*}$ on $L \cup T' \cup T$.
- Return model $f_{\theta}^{\lambda^*}$ and estimate $\hat{R}_T(f_{\theta'}^{\lambda^*})$.



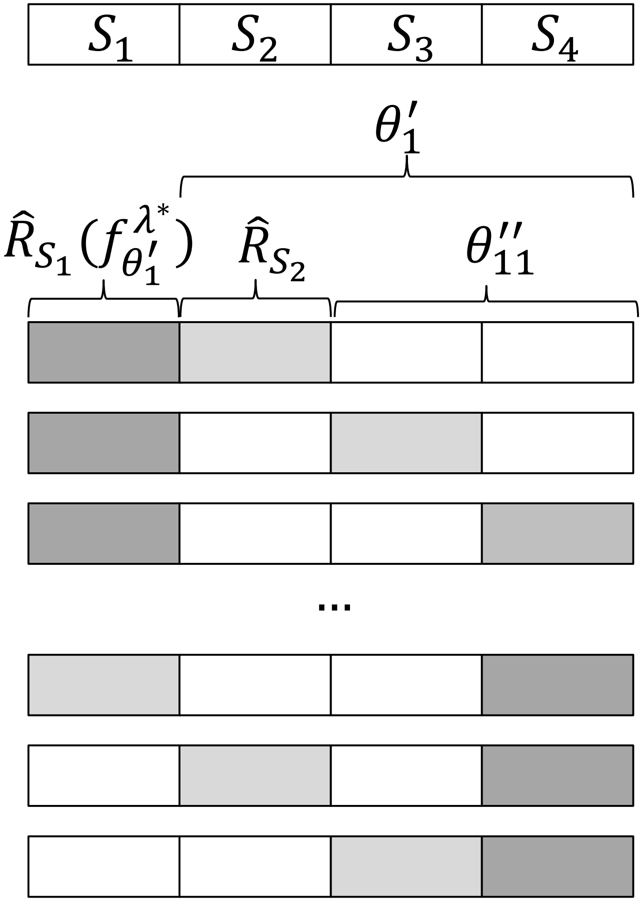
Triple Cross Validation: Analysis

- Empirical risk $\hat{R}_T(\theta')$ is a pessimistic estimator for$R(\theta)$ because θ' is trained on less data than θ.
- λ^* may be a poor estimate of the optimal parameters because T' may be small.
- The variance of $\hat{R}_T(\theta')$ may high because T may be small.
- Protocol is used when the total sample S is very large.



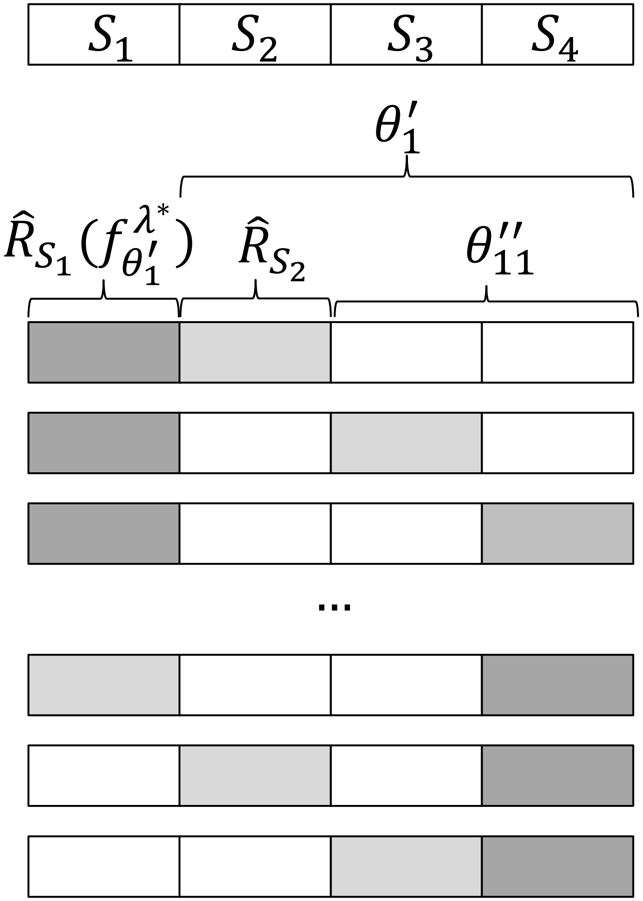
Nested Cross Validation

- For $i = 1 \ldots k$
  - Iterate over values λ
    - For $j = 1 \ldots k \setminus i$
      - Train $f_{\theta_{ij}}^{\lambda}$ on $S \setminus S_i \setminus S_j$
      - Determine $\hat{R}_{S_j}\left(f_{\theta_{ij}}^{\lambda}\right)$
    - Average $\hat{R}_{S_j}$ to determine $\hat{R}_{S \setminus S_i}\left(f_{\theta_i'}^{\lambda}\right)$
  - Choose λ_i^* that minimizes $\hat{R}_{S \setminus S_i}\left(f_{\theta_i'}^{\lambda}\right)$
  - Train $f_{\theta_i}^{\lambda_i^*}$ on $S \setminus S_i$
  - Determine $\hat{R}_{S_i}\left(f_{\theta_i'}^{\lambda_i^*}\right)$
- Average $\hat{R}_{S_i}\left(f_{\theta_i}^{\lambda_i^*}\right)$ to determine $\hat{R}_S(f_{\theta^*}^{\lambda^*})$
- Determine λ^* by simple cross validation (iterate over values λ, train on $S \setminus S_i$, measure $\hat{R}_{S_i}(f_{\theta_i}^{\lambda})$)
- Train $f_\theta^{\lambda^*}$ on S
- Return $f_\theta^{\lambda^*}$ and $\hat{R}_S(f_{\theta^*}^{\lambda^*})$



Nested Cross Validation: Analysis

- Complextiy: k^2 models have to be trained and evaluated
- Slightly pessimistic because $f_\theta^{\lambda^*}$ has been trained on more data than the $f_{\theta_i}^{\lambda_i^*}$.
- Lower variance than triple cross validation because all data is used for evaluation
- Better estimate of λ^* because almost all data is used for tuning.
- Best tuning protocol when few data are available.



Summary

- Risk: expected loss over input distribution $p(\mathbf{x}, y)$.
- Empirical risk: estimate of risk on data.
- Precision-recall curves and ROC curves characterize decision function. Each point on curve is classifier for some threshold θ_0.
- Evaluation protocols:
  - Hold-out testing: good for large samples
  - K-fold Cross Validation: good for small samples.
- Model selection: tune model hyperparameters.
  - Triple cross validation: good for large samples.
  - Nested cross validation: good for small samples.