Introduction to Statistical Machine Learning

Christfried Webers

Statistical Machine Learning Group
NICTA
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2013

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

ISML 2013

Outlines

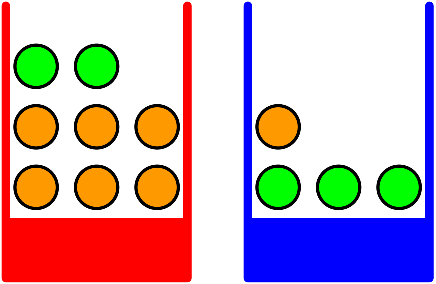
Part IV

Probability and Uncertainty

Simple Experiment

1. Choose a box.
   - Red box $p(B = r) = 4/10$
   - Blue box $p(B = b) = 6/10$
2. Choose any item of the selected box with equal probability.

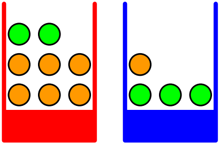
- Given that we have chosen an orange, what is the probability that the box we chose was the blue one?



What do we know?

- $p(F = o \mid B = b) = 1/4$
- $p(F = a \mid B = b) = 3/4$
- $p(F = o \mid B = r) = 3/4$
- $p(F = a \mid B = r) = 1/4$
- Given that we have chosen an orange, what is the probability that the box we chose was the blue one?

$$p(B = b \mid F = o)?$$



NICTA
The Australian National University

Calculating the Posterior $p(B = b \,|\, F = o)$

- Bayes' Theorem

$$p(B = b \,|\, F = o) = \frac{p(F = o|B = b)p(B = b)}{p(F = o)}$$

- Sum Rule for the denominator

$$\begin{aligned} p(F = o) &= p(F = o, B = b) + p(F = o, B = r) \\ &= p(F = o|B = b)p(B = b) \\ &\quad + p(F = o|B = r)p(B = r) \\ &= \frac{1}{4} \times \frac{6}{10} + \frac{3}{4} \times \frac{4}{10} = \frac{9}{20} \end{aligned}$$

- $$p(B = b \,|\, F = o) = \frac{1}{4} \times \frac{6}{10} \times \frac{20}{9} = \frac{1}{3}$$

Bayes' Theorem - revisited

- Before choosing an item from a box: most complete information in $p(B)$ (prior).
- Note, that in our example $p(B = b) = \frac{6}{10}$. Therefore choosing the box from the prior, we would opt for the blue box.
- Once we observe some data (e.g. choose an orange), we can calculate $p(B = b \mid F = o)$ (posterior probability) via Bayes' theorem.
- After observing an orange, the posterior probability $p(B = b \mid F = o) = \frac{1}{3}$ and therefore $p(B = r \mid F = o) = \frac{2}{3}$.
- Observing an orange it is now more likely that the orange came from the red box.

Bayes' Rule

$$\overbrace{p(Y \mid X)}^{\text{posterior}} = \frac{\overbrace{p(X \mid Y)}^{\text{likelihood}} \; \overbrace{p(Y)}^{\text{prior}}}{\underbrace{p(X)}_{\textit{normalisation}}} = \frac{p(X \mid Y)\, p(Y)}{\sum_Y p(X \mid Y) p(Y)}$$

Bayes' Probabilities

- classical or frequentist interpretation of probabilities
- Bayesian view: probabilities represent uncertainty
- Example: Will the Arctic ice cap have disappeared by the end of the century?
- fresh evidence can change the opinion on ice loss
- goal: quantify uncertainty and revise uncertainty in light of new evidence
- use Bayesian interpretation of probability

Andrey Kolmogorov - Axiomatic Probability Theory (1933)

- Let (Ω, F, P) be a measure space with $P(\Omega) = 1$.
- Then (Ω, F, P) is a probability space, with sample space Ω, event space F and probability measure P.
- 1. Axiom $P(E) \geq 0 \quad \forall E \in F$.
- 2. Axiom $P(\Omega) = 1$.
- 3. Axiom $P(E_1 \cup E_2 \cup \dots) = \sum_i P(E_i)$ for any countable sequence of pairwise disjoint events $E_1, E_2, \dots$.

Richard Threlkeld Cox - (1946, 1961)

- Assume numerical values are used to represent degrees of belief.
- Define a set of axioms encoding common sense properties of such beliefs.
- Results in a set of rules for manipulating degrees of belief which are equivalent to the sum and product rule of probability.
- many other authors have proposed different sets of axioms and properties
- Result: the numerical quantities behave all according to the rules of probability
- Denote these quantities as Bayesian probabilities.

Curve fitting - revisited

- uncertainty about the parameter $\mathbf{w}$ captured in the prior probability $p(\mathbf{w})$
- observed data $\mathcal{D} = \{t_1, \dots, t_N\}$
- calculate the uncertainty in $\mathbf{w}$ after the data $\mathcal{D}$ have been observed

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w}) p(\mathbf{w})}{p(\mathcal{D})}$$

- $p(\mathcal{D} \mid \mathbf{w})$ as a function of $\mathbf{w}$: likelihood function
- likelihood expresses how probable the data are for different values of $\mathbf{w}$
- not a probability function over $\mathbf{w}$

Likelihood Function - Frequentist versus Bayesian

Likelihood function $p(\mathcal{D} \mid \mathbf{w})$

Frequentist Approach

- $\mathbf{w}$ considered fixed parameter
- value defined by some 'estimator'
- error bars on the estimated $\mathbf{w}$ obtained from the distribution of possible data sets $\mathcal{D}$

Bayesian Approach

- only one single data set $\mathcal{D}$
- uncertainty in the parameters comes from a probability distribution over $\mathbf{w}$

a distribution from Bayesians

one single value from frequentist

Frequentist Estimator - Maximum Likelihood

- choose $\mathbf{w}$ for which the likelihood $p(\mathcal{D} \,|\, \mathbf{w})$ is maximal
- choose $\mathbf{w}$ for which the probability of the observed data is maximal
- Machine Learning: error function is negative log of likelihood function
- log is a monoton function
- maximising the likelihood $\iff$ minimising the error
- Example: Fair-looking coin is tossed three times, always landing on heads.
- Maximum likelihood estimate of the probability of landing heads will give 1.

Bayesian Approach

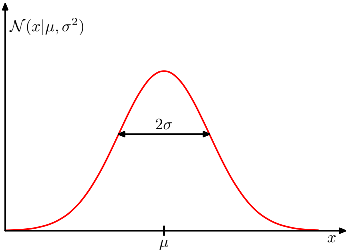
- including prior knowledge easy (via prior $\mathbf{w}$)
- BUT: if prior is badly chosen, can lead to bad results
- subjective choice of prior
- sometimes choice of prior motivated by convinient mathematical form
- need to sum/integrate over the whole parameter space
- advances in sampling (Markov Chain Monte Carlo methods)
- advances in approximation schemes (Variational Bayes, Expectation Propagation)

In Murphy's book, he use Bernouli distribution to denote prior of binomial likelihood.

The Gaussian Distribution

- $x \in \mathbb{R}$
- Gaussian Distribution with mean μ and variance σ^2

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$



The Gaussian Distribution

- $\mathcal{N}(x \,|\, \mu, \sigma^2) > 0$
- $\int_{-\infty}^{\infty} \mathcal{N}(x \,|\, \mu, \sigma^2) \, \mathrm{d}x = 1$
- Expectation over x

$$\mathbb{E}\left[x\right] = \int_{-\infty}^{\infty} \mathcal{N}(x \,|\, \mu, \sigma^2) \, x \;\; \mathrm{d}x = \mu$$

- Expectation over x^2

$$\mathbb{E}\left[x^2\right] = \int_{-\infty}^{\infty} \mathcal{N}(x \,|\, \mu, \sigma^2) \, x^2 \;\; \mathrm{d}x = \mu^2 + \sigma^2$$

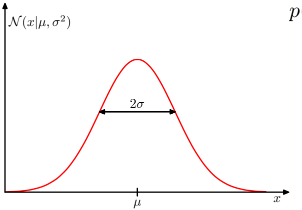
- Variance of x

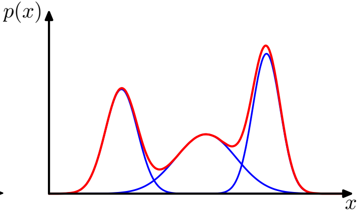
$$\mathrm{var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2 = \sigma^2$$

Christfried Webers
NICTA
The Australian National University

The Mode of a Probability Distribution

- Mode of a distribution : the value that occurs the most frequently in a probability distribution.
- For a probability density function: the value x at which the probability density attains its maximum.
- Gaussian Distribution has one mode (unimodular); the mode is μ.
- If there are multiple local maxima in the probability distribution, the probability distribution is called multimodal (example: mixture of three Gaussians).





The Bernoulli Distribution

- Two possible outcomes $x \in \{0, 1\}$ (e.g. coin which may be damaged).
- $p(x = 1 \mid \mu) = \mu$ for $0 \leq \mu \leq 1$
- $p(x = 0 \mid \mu) = 1 - \mu$
- Bernoulli Distribution

$$\mathrm{Bern}(x \mid \mu) = \mu^{x}(1 - \mu)^{1-x}$$

- Expectation over x

$$\mathbb{E}\left[x\right] = \mu$$

- Variance of x

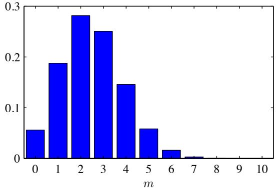
$$\mathrm{var}[x] = \mu(1 - \mu)$$

Christfried Webers
NICTA
The Australian National University

The Binomial Distribution

- Flip a coin N times. What is the distribution to observe heads exactly m times?
- This is a distribution over $m = \{0, \dots, N\}$.
- Binomial Distribution

$$\mathrm{Bin}(m \,|\, N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

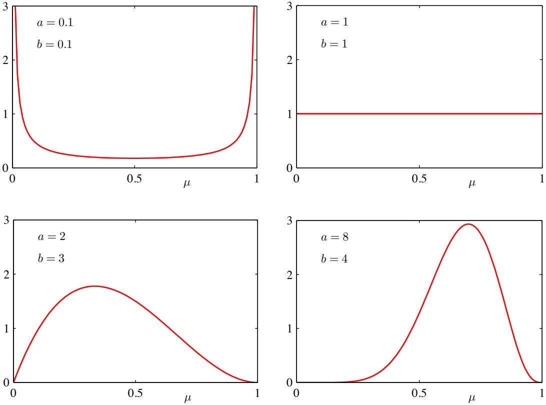


$$N = 10, \mu = 0.25$$

The Beta Distribution

- Beta Distribution

$$\mathrm{Beta}(\mu \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$$

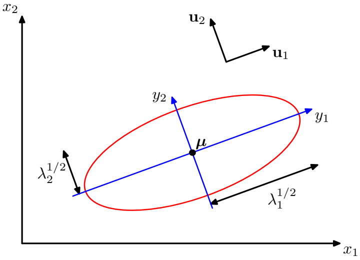


The Gaussian Distribution over a Vector $\mathbf{x}$

- $\mathbf{x} \in \mathbb{R}^D$
- Gaussian Distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$

$$\mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\}$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.



The Gaussian Distribution over a vector $\mathbf{x}$

- Can find a linear transformation to a new coordinate system $\mathbf{y}$ in which the $\mathbf{x}$ becomes

$$\mathbf{y} = \mathbf{U}^T \; (\mathbf{x} - \boldsymbol{\mu}) \, ,$$

- $\mathbf{U}$ is the eigenvector matrix for the covariance matrix $\boldsymbol{\Sigma}$ with eigenvalue matrix $\mathbf{E}$

$$\boldsymbol{\Sigma}\,\mathbf{U} = \mathbf{U}\,\mathbf{E} = \mathbf{U}\begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_D \end{bmatrix}$$

- $\mathbf{U}$ can be made an orthogonal matrix, therefore the columns u_i of $\mathbf{U}$ are unit vectors which are orthogonal to each other $u_i^T u_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$
- Now we can write $\boldsymbol{\Sigma}$ and its inverse (prove that $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} = \mathbf{I}$)

$$\boldsymbol{\Sigma} = \mathbf{U}\,\mathbf{E}\,\mathbf{U}^T = \sum_{i=1}^{n} \lambda_i \, u_i \, u_i^T \qquad \boldsymbol{\Sigma}^{-1} = \mathbf{U}\,\mathbf{E}^{-1}\,\mathbf{U}^T = \sum_{i=1}^{n} \frac{1}{\lambda_i} \, u_i \, u_i^T$$

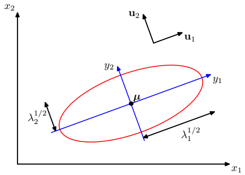
The Gaussian Distribution over a vector $\mathbf{x}$

- Now use the linear transformation $\mathbf{y} = \mathbf{U}^T\ (\mathbf{x} - \boldsymbol{\mu})$ and $\boldsymbol{\Sigma}^{-1} = \mathbf{U}\,\mathbf{E}^{-1}\,\mathbf{U}^T$ to transform the exponent $(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ into

$$\mathbf{y}^T\,\mathbf{E}\,\mathbf{y} = \sum_{j=1}^{n} \frac{y_j^2}{\lambda_j}$$

- Now exponating the sum (and taking care of the factors) results in a product of scalar valued Gaussian distributions in orthogonal directions u_i

$$p(\mathbf{y}) = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\{-\frac{y_j^2}{2\lambda_j}\}$$



Partitioned Gaussians

- Given a joint Gaussian distribution $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Sigma}$ the covariance matrix, and $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ the precision matrix
- Assume that the variables can be partitioned into two sets

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \ \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

$$\begin{aligned} \boldsymbol{\Lambda}_{aa} &= (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1} \\ \boldsymbol{\Lambda}_{ab} &= -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1} \end{aligned}$$

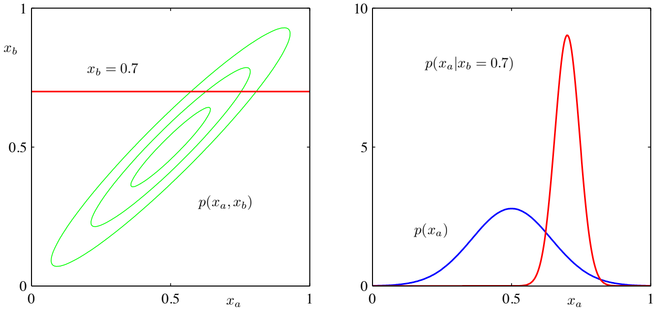
- Conditional distribution

$$\begin{aligned} p(\mathbf{x}_a \mid \mathbf{x}_b) &= \mathcal{N}(\mathbf{x}_a \mid \boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

- Marginal distribution

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a \mid \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

Partitioned Gaussians



Contours of a Gaussian distribution over two variables x_a and x_b (left), and marginal distribution $p(x_a)$ and conditional distribution $p(x_a \mid x_b)$ for $x_b = 0.7$ (right).

Nonlinear Change of Variables in Distributions

- Given some $p_x(x)$.
- Consider a nonlinear change of variables

$$x = g(y)$$

- What is the new probability distribution $p_y(y)$ in terms of the variable y ?

$$\begin{aligned} p_y(y) &= p_x(x) \left| \frac{dx}{dy} \right| \\ &= p_x(g(y)) \left| g'(y) \right| \end{aligned}$$

- For vector valued $\mathbf{x}$ and $\mathbf{y}$

$$p_y(\mathbf{y}) = p_x(\mathbf{x}) \; |\, \mathbf{J} \,|$$

where $J_{ij} = \frac{\partial x_i}{\partial y_j}$.

Decision Theory - Key Ideas

NICTA
The Australian National University

- Two classes $\mathcal{C}_1$ and $\mathcal{C}_2$
- joint distribution $p(\mathbf{x}, \mathcal{C}_k)$
- using Bayes' theorem

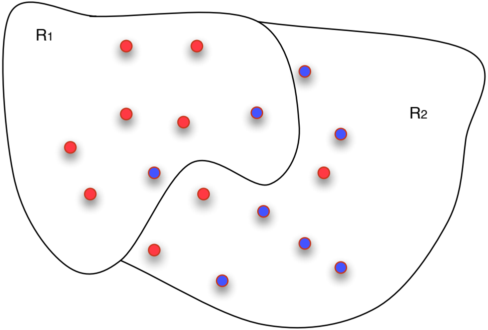
$$p(\mathcal{C}_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathcal{C}_k)\, p(\mathcal{C}_k)}{p(\mathbf{x})}$$

- Example: cancer treatment ($k = 2$)
- data $\mathbf{x}$: an X-ray image
- $\mathcal{C}_1$: patient has cancer ($\mathcal{C}_1$: patient has no cancer)
- $p(\mathcal{C}_1)$ is the prior probability of a person having cancer
- $p(\mathcal{C}_1 \mid \mathbf{x})$ is the posterior probability of a person having cancer after having seen the X-ray data

Decision Theory - Key Ideas

- Need a rule which assigns each value of the input $\mathbf{x}$ to one of the available classes.
- The input space is partitioned into decision regions $\mathcal{R}_k$.
- Leads to decision boundaries or decision surfaces
- probability of a mistake

$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}
\end{aligned}
$$

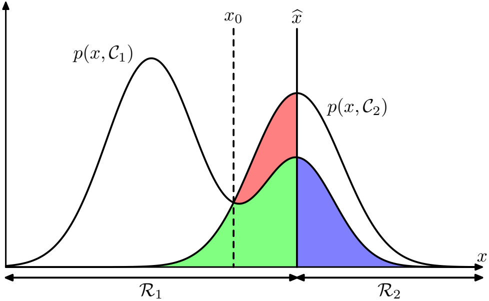


Decision Theory - Key Ideas

- probability of a mistake

$$p(\text{mistake}) = p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1)$$
$$= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\, \mathrm{d}\mathbf{x}$$

- goal: minimize $p(\text{mistake})$



Decision Theory - Key Ideas

- multiple classes
- instead of minimising the probability of mistakes, maximise the probability of correct classification

$$p(\text{correct}) = \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k)$$

$$= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}$$

Minimising the Expected Loss

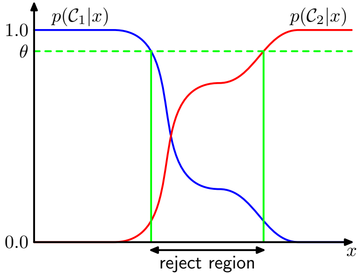
- Not all mistakes are equally costly.
- Weight each misclassification of $\mathbf{x}$ to the wrong class $\mathcal{C}_j$ instead of assigning it to the correct class $\mathcal{C}_k$ by a factor L_{kj}.
- The expected loss is now

$$\boxed{\mathbb{E}\left[L\right] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj}\, p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}}$$

- Goal: minimize the expected loss $\mathbb{E}\left[L\right]$

The Reject Region

- Avoid making automated decisions on difficult cases.
- Difficult cases:
  - posterior probabilities $p(\mathcal{C}_k \mid \mathbf{x})$ are very small
  - joint distributions $p(\mathbf{x}, \mathcal{C}_k)$ have comparable values



S-fold Cross Validation

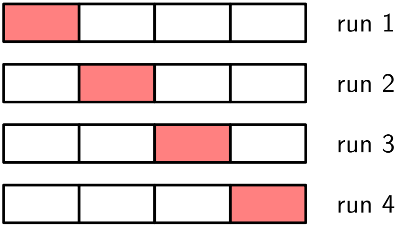
- Given a set of N data items and targets.
- Goal: Find the best model (type of model, number of parameters like the order p of the polynomial or the regularisation constant λ). Avoid overfitting.
- Solution: Train a machine learning algorithm with some of the data, evaluate it with the rest.
- If we have many data
  1. Train a range of models or a model with a range of parameters.
  2. Compare the performance on an independent data set (validation set) and choose the one with the best predictive performance.
  3. Still, overfitting to the validation set can occur. Therefore, use a third test set for final evaluation. (Keep the test set in a safe and never give it to the developers ;–)

S-fold Cross Validation

- For few data, there is a dilemma: Few training data or few test data.
- Solution is cross-validation.
- Use a portion of $(S-1)/S$ of the available data for training, but use all the data to asses the performance.
- For very scarce data one may use $S = N$, which is also called the leave-one-out technique.

S-fold Cross Validation

- Partition data into S groups.
- Use $S - 1$ groups to train a set of models that are then evaluated on the remaining group.
- Repeat for all S choices of the held-out group, and average the performance scores from the S runs.



Example for $S = 4$.