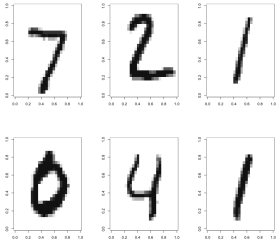
Example: classifying hadwritten digits



Example: classifying hadwritten digits

Training data: 60000 images.
Test data: 10000 images.
Features: intensities (0-255, scaled to 0-1) in the $28 \times 28 = 784$ pixels as features.

Methods:
- Multinomial regression with LASSO prior
- Support vector machines
- Neural Networks (deep?)

Example: classifying hadwritten digits

- Confusion matrix

PREDICTION / TRUTH

	0	1	2	3	4	5	6	7	8	9
0	966	0	8	1	1	7	9	2	4	6
1	0	1121	1	1	0	2	3	13	7	7
2	2	2	957	13	5	4	4	21	7	0
3	0	2	9	947	0	29	1	3	12	10
4	0	0	12	1	940	5	5	9	8	32
5	6	1	3	19	1	816	9	1	24	9
6	4	4	13	1	7	12	926	0	10	1
7	1	0	9	10	2	2	0	954	5	13
8	1	4	17	11	2	10	1	3	892	4
9	0	1	3	6	24	5	0	22	5	927

Example: smartfone typing predictions

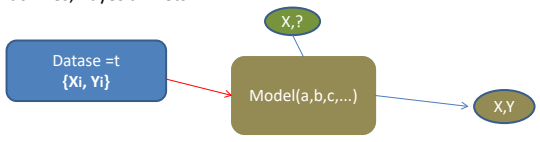
Example: smartfone typing predictions

- Assume a simple (Markov) model of a sentence:
$$p(w_1, \dots, w_n) = p(w_1)p(w_2|w_1) \dots p(w_n|w_{n-1})$$
- Intuition:
  - $p(person|crazy) = 0.1$
  - $p(horse|crazy) = 0.0001$

  Highest P(?|Donald) ?
- Probability for sentence depends only on $p(w_n|w_{n-1})$
- How to compute ? Investigate a lot of data!
$$p(w_k|w_{k-1}) = \frac{\#\ cases\ w_k\ follows\ w_{k-1}}{\#\ cases\ w_k}$$
- In practice, more advanced model used
  - Neural networks for ex.

Types of learning

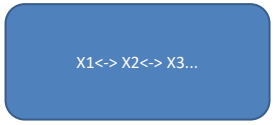
- **Supervised learning** (classification, regression)
  - Compute parameters from data
  - Given features of a new object, predict target
  - **Classification** (Y=categorical), **Regression** (Y=continuous)
- Most of ML models: Neural Nets, Decision Trees, Support Vector Machines, Bayesian nets



Types of learning

- Unsupervised learning (→Data Mining)
  - No target
  - Aim is to extract interesting information about
    - Relations of parameters to each other
    - Grouping of objects

Ex: clustering, density estimation, association analysis



Types of learning

- **Semi-supervised**: targets are known only for some observations.
- **Active learning**. Strategies for deciding which observations to label
- **Reinforcement learning**. Find suitable actions to maximize the reward. True targets are discovered by trial and error.

Basic ML ingridients

- **Data** D: observations (cases)
  - Features $X_1, \dots X_p$
  - Targets $Y_1, \dots, Y_r$

Case	X_1	X_2	Y
1			
2			
...			

- **Model** $P(x|\ w_1, \dots w_k)$ or $P(y|x, w_1, \dots w_k)$
  - Example: Linear regression $p(y|x, w) = N(w_0 + w_1 x, \sigma^2)$
- **Learning procedure** (data→get parameters $\hat{w}$ or $p(w|D)$)
  - Maximum likelihood, Bayesian estimation…
- **Prediction** of new data X^{new} by using the fitted model

Types of data sets

- **Training data** (training set D): used for fitting the model
  - Supervised learning: w_i in $P(y|\boldsymbol{x}, w_1, \dots w_k)$ estimated using D

X	Y
1.1	M
2.3	F

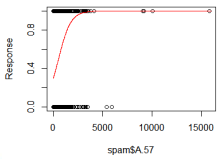
- **Test data** (test set T): used for predictions
  - Supervised learning: estimate $p(Y)$ or $\hat{Y}$ for new $\boldsymbol{x}$

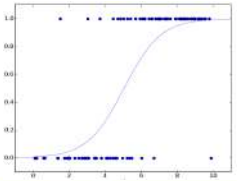
X	Y
1.3	?
2.9	?

Logistic regression

- Data $Y_i \in \{Spam, Not\ Spam\}, X_i = \#of\ a\ word$
- Model: $p(Y = Spam|w, x) = \frac{1}{1+e^{-w_0-w_1 X}}$
- Fitting: maximum likelihood
- Prediction : $p(spam) = p(Y = spam|x)$

We can also make point predictions -how?

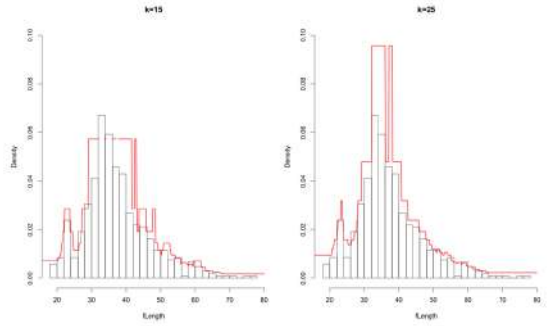




K-nearest neighbor density estimation

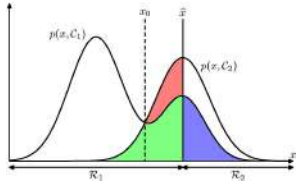
- Data: Fish length $X_1, \dots X_N$
- Model $p(x|K) = \frac{K}{N \cdot \Delta}$
  - K: #neighbors in training data
  - Δ: length of the interval containing K neighbors
- Learning: Fix some K or find an appropriate K
- Prediction: predict $p(x|K)$

K-nearest neighbor density estimation



K-nearest neighbor density estimation

- Why estimating a density can be interesting:
  1. Estimate **class-conditional densities** $p(x|y = C_i)$
  2. Predict

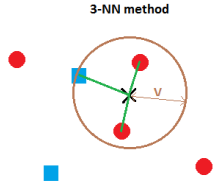


K-nearest neighbor classification

- Given N observations $(\boldsymbol{X}_j, Y_j)$
  - $Y_j = C_i$, where $C_1, \dots C_m$ are possible class values
- Model assumptions
  - Apply K-NN density estimation:

$$p(X = x|Y = C_i) = \frac{K_i}{N_i V}, p(C_i) = \frac{N_i}{N}$$

  - V: volume of the sphere
  - K_i: #obs from training data of $Y = C_i$ in the sphere
  - N_i: #obs from training data of $Y = C_i$



Bayesian classification

- Prediction $\hat{Y}(\boldsymbol{x}) = C_l$

$$l = \arg \max_{i \in \{1,\dots,m\}} p(C_i|\boldsymbol{x})$$

- Bayes theorem

$$p(C_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_i)p(C_i)}{p(\boldsymbol{x})}$$

- We get

$$p(C_i|x) \propto \frac{K_i}{K}$$

K-nearest neighbor classification

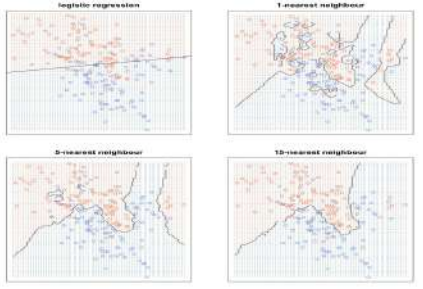
Algorithm

1. Given training set D,number K, and test set T
2. For each $x \in T$
   1. For each $i = 1, \dots M$
      1. $p'(C_i|x) = \frac{K_i}{K}$
   2. Compute $l = \arg \max_{i \in \{1,\dots,m\}} p'(C_i|\boldsymbol{x})$
   3. Predict $\hat{Y}(x) = C_l$

Majority voting: prediction for x is defined by majority voting of K neighbors

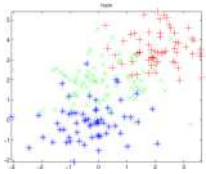
K-nearest neigbor example

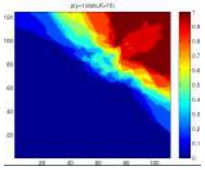
Why classification results are so different for K-NN?



Model types

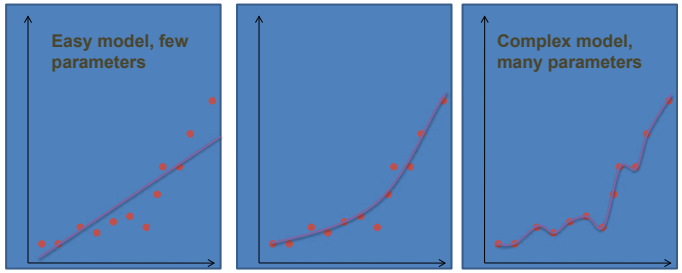
- Parametric models
  - Have certain number of parameters independently of the size of training data
  - Assumption about of the data distribution
  - Ex: logistic regression
- Nonparametric models
  - Number of parameters (complexity) grows with training data
    - Example: K-NN classifier





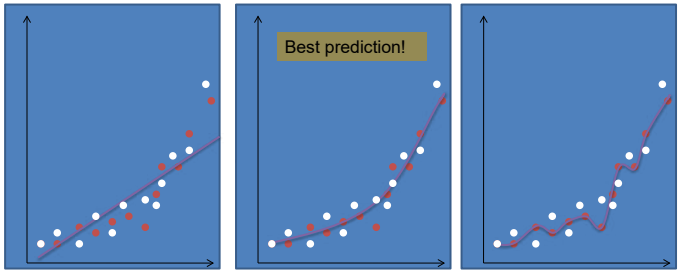
Overfitting

- Which model feels appropriate?



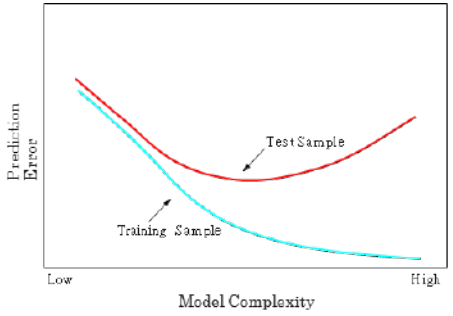
Overfitting

Now new data from the same process



Overfitting

- Observed:



Model selection

- Given several models $M_1, \dots M_m$
- Divide data set into **training** and **test** data

Training	Test

- Fit models M_i to training data→get parameter values
- Use fitted models to predict test data and compare **test errors** $R(M_1), \dots R(M_m)$
- Model with lowest prediction error is best

Comment:

- Approach works well for moderate/large data

Typical error functions

- Regression, **MSE** :

$$R(Y, \hat{Y}) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

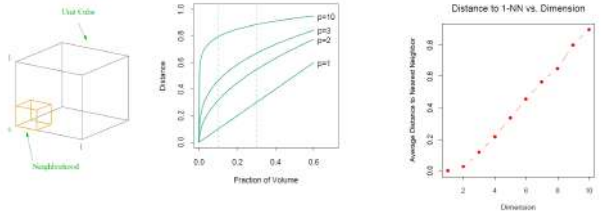
- Classification, **misclassification rate**

$$R(Y, \hat{Y}) = \frac{1}{N}\sum_{i=1}^{N} I(Y_i \neq \hat{Y}_i)$$

Curse of dimensionality

- Given data D:
  - Features $X_1, \dots X_p$
  - Targets $Y_1, \dots, Y_r$
- When p increases models using ”proximity” measures work badly
- Curse of dimensionality: A point has no ”near neighbors” in high dimensions → using class labels of a neighbor can be misleadning
  - Distance-based methods affected

Curse of dimensionality



Curse of dimensionality

- Hopeless? No!
- Real data normally has much lower effective dimension
  - Dimensionality reduction techniques
- Smoothness assumption
  - small change in one of Xs should lead to small change in Y→interpolation

Probability

How likely it is that some event will happen?

Idea:

- Experiment
- Outcomes (sample points) O_1, O_2,… O_n
- Sample space Ω
- Event A
- Probability function P: Events →[0,1]

Probability

Example: Tossing a coin two times

Example:

- $p(A)$ frequency of observing A
- $p(A,B)$ frequency of observing A and B
- $p(B|A)$ frequency of observing B given A

Properties and definitions

- One can think of events as sets
  - Set operations are defined: $A \cup B, A \cap B, \bar{A} \backslash B$
- $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$
- Independence $P(A,B) \equiv P(A \cap B) = P(A)P(B)$
- Conditional probability $P(A|B) = \frac{P(A,B)}{P(B)}$

Bayes theorem

Example:

- We have constructed spam filter that
  - identifies spam mail as spam with probability 0.95
  - Identifies usual mail as spam with probability 0.005
- This kind of spam occurs once in 100,000 mails
- If we found that a letter is a spam, what is the probability that it is actually a spam?

Bayes theorem

- We have some knowledge about event B
  - Prior probability P(B) of B
- We get new information A
  - P(A)
  - P(A|B) probability of A can occur given B has occured
- New (updated) knowledge about B
  - Posterior probability P(B|A)

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Random variables

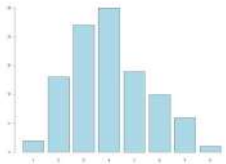
- Instead of having events, we can have a variable X:
  - Events→$\mathbb{R}$ Continuous random variables
  - Events→$\mathbb{N}$ Discrete random variables

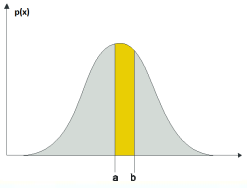
Examples:

- X={amount of times the word "crisis" can be found in financial documents}
  - P(X=3)
- X={Time to download a specific file to a specific computer}
  - P(X=0.36 min)

Distributions

- Discrete
  - Probability mass function P(x) for all feasible x
- Coninuous
  - Probability density function p(x)
    - $p(x \in [a,b]) = \int_a^b p(x)dx$
    - $p(x) \geq 0, \int_{-\infty}^{+\infty} p(x)dx = 1$
  - Cumulative distribution function $F(x) = \int_0^x p(t)dt$





Expected value and variance

- Expected value = mean value
  - $E(X) = \sum_{i=1}^{n} X_i P(X_i)$
  - $E(X) = \int Xp(X)dX$
- Variance how much values of random variable can deviate from mean value
  - $Var(X) = E\left(X - E(X)\right)^2 = E(X^2) - E(X)^2$

Probabilities

- **Laws of probabilities**
  - Sum rule (compute **marginal** probability)

$$p(X) = \sum_Y p(X,Y)$$

$$p(X) = \int p(X,Y)dY$$

  - Product rule

$$p(X,Y) = p(X|Y)p(Y)$$

Combination 1:

$$p(X) = \sum_Y p(X|Y)p(Y)$$

$$p(X) = \int p(X|Y)p(Y)dY$$

Bayes theorem

For random variables:

Bayes Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(Y|X) \propto p(X|Y)p(Y)$$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{\int p(X|Y)p(Y)dY}$$

Some conventional distributions

Bernoulli distribution

- Events: Success (X=1) and Failure (X=0)
- P(X=1)=p, P(X=0)=1-p

- $E(X) = p$
- $Var(X) = 1 - p$

Examples: Tossing coin, vinning a lottery,..

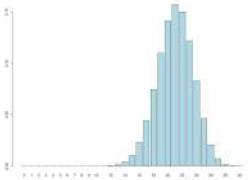
Some conventional distributions

Binomial distribution

- Sequence of *n* Bernoulli events
- X={Amount of successes among these events}, X=0,…,n

$$P(X = r) = \frac{n!}{(n-r)!\,r!} p^r (1-p)^{n-r}$$

- $EX = np$
- $Var(X) = np(1-p)$



Poisson distribution

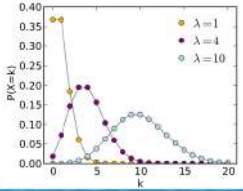
- Customers of a bank **n** (in theory, endless population)
- Probability that a specific person will make a call to the bank between 13.00 and 14.00 a certain day is ***p***
  - ***p*** can be very small if population is large (rare event)
  - Still, some people will make calls between 13.00 and 14.00 that day, and their amount may be quite big
  - A known quantity **λ=*np*** is mean amount of persons that call between 13.00 and 14.00
  - **X**={amount of persons that have called between 13.00 and 14.00}

Poisson distribution

- $P(X = r) = \lim_{n\to\infty} \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r}$
- It can be shown that

$$P(X = r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

- $E(X) = \lambda$
- $Var(X) = \lambda$



Poisson distribution

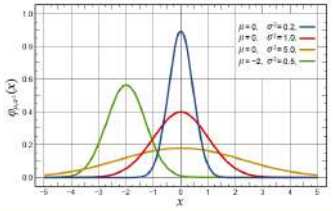
- Further properties:
  - Poisson distribution is a good approximation of the binomial distribution if n >20 and $p < 0.05$
  - Excellent approximation if $n \geq 100$ and $np \leq 10$

Normal distribution

- Appears in almost all applications
  - Difference between the times required to download two specific documents to a specific computer

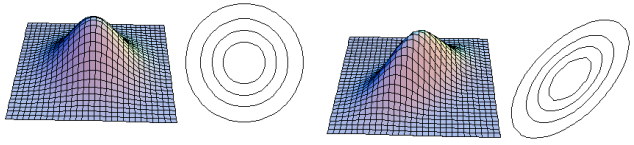
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \sigma > 0$$

- $E(X) = \mu$
- $Var(X) = \sigma^2$



Multivariate distributions

- Probability of two variables having certain values at the same time
  - P.D.F. p(x,y)
  - Correlation



Basic ML ingridients

- Data D: observations
  - Features $X_1, \ldots X_p$
  - Targets $Y_1, \ldots, Y_r$

Case	X_1	X_2	Y
1			
2			
…			

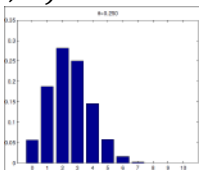
- Model $P(x|\, w_1, \ldots w_k)$ or $P(y|x, w_1, \ldots w_k)$
  - Example: Linear regression $p(y|x,w) = N(w_0 + w_1 x, \sigma^2)$
- Learning procedure (data→get parameters $\hat{w}$ or $p(w|D)$)
  - Maximum likelihood, Bayesian estimation
- Predict new data X^{new} by using the fitted model

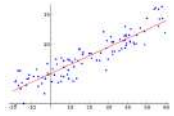
Probabilistic models

- A distribution $p(x|w)$ or $p(y|x,w)$
- Example:
  - $x \sim Bin(n,\theta)$

    $p(x=k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$

  - $y \sim N(\alpha_0 + \alpha_1 x, \sigma^2)$





Learn basic distributions and their properties→PRML, chapter 2!

Fitting a model

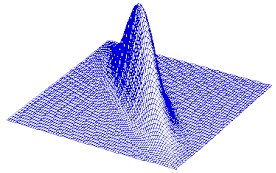
- Given dataset D and model $p(x|w)$ or $p(y|x,w)$
  - Frequentist approach: which combination of parameter values fits my data best?
  - Bayesian approach: parameters are random variables, all feasible values are acceptable
    - Different parameter values have different probabilities

Fitting a model

- Frequenist principle: **Maximum likelihood** principle
  - Compute likelihood $p(\boldsymbol{D}|\, w)$

$$p(\boldsymbol{D}|\, w) = \prod_{i=1}^{n} p(X_i|w)$$

$$p(\boldsymbol{D}|\, w) = \prod_{i=1}^{n} p(Y_i|X_i, w)$$



  - Maximize the likelihood and find the optimal w^*

Fitting a model

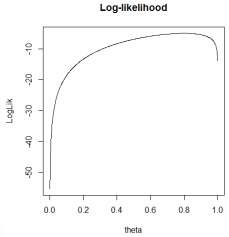
Remarks:

- Likelihood shows how much the chosen parameter value is proper for a specific model and the given data
- Normally **log-likelihood** is used in computations instead
- Other alternatives to ML exist…

Fitting a model

Example: tossing a coin.

$D = \{0,1,1,0,1,1,1,1,1,1,1,1\}$,
$p(x=1|\theta) = \theta, p(x=0|\theta) = 1-\theta$



Bayesian probabilities

- Probability reflects your knowledge (uncertainty) about a phenomenon → **subjective probabilities**
  - **Prior probability** $p(w)$, can be uninformative $p(w) \propto 1$
  - Formulate a model, compute **likelihood** $p(D|w)$
  - **Posterior probability** $p(w|D)$, after observing data
    - $p(w|D) \propto p(D|w)p(w)$
- Model parameters are considered as random variables
  - In real life, do not need to be random, but we model as random

Fitting a model

- Bayesian principle
  - Compute $p(w|D)$and then decide yourself what to do with this (for ex. MAP, mean, median)
- Use bayes theorem

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} \propto p(D|w)p(w)$$

- $p(D)$ is **marginal likelihood**
  - $p(D) = \int p(D|w)p(w)dw$ or
  - $p(D) = \sum_i p(D|w_i)p(w_i)$

Example: tossing a coin. Find $p(\theta|D)$, estimate posterior mean θ^*

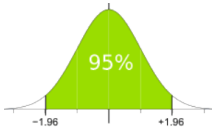
Fitting a model

- How to chose the prior?
  - Expert knowledge about the phenomenon
  - Forcing a model to have a certain structure
    - Example: decision trees: prior prefers smaller trees

      http://en.wikipedia.org/wiki/Conjugate_prior
  - Conjugacy
    - Distribution of the posterior is the same type as the distribution of the likelihood or prior
- Prior is the most controversial about Bayesian methods, but
  - When $N \to \infty$, data overwhelms the prior

Measuring uncertainty

- **Confidence interval** (frequentist)
  1. Model $p(x|w)$ is known
  2. $\hat{w}$ is a function of x by ML
  3. Derive distribution of $\hat{w}$
  4. Compute quantiles



- **Credible interval** (Bayes)
- **Prediction interval** (models)

- Example: Prediction interval for $Y \sim N(2x+4, 1)$ at $x = 5$

Regression and regularization

Lecture 1d

Overview

- Linear regression
- Ridge Regression
- Lasso
- Variable selection

Simple linear regression

Model:

$$y \sim N(w_0 + w_1 x, \sigma^2)$$

or

$$y = w_0 + w_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

or

$$p(y|x, w) = N(w_0 + w_1 x, \sigma^2)$$

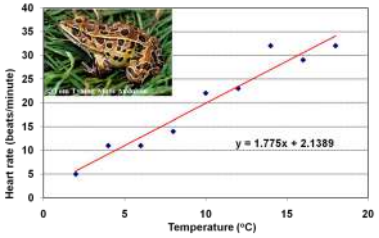
Terminology:

w_0: **intercept (or bias)**

w_1: **regression coefficient**

Response

The target responds directly and linearly to changes in the feature



Ordinary least squares regression (OLS)

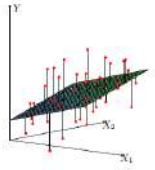
Model:

$$y \sim N(\boldsymbol{w}^T \boldsymbol{x}, \sigma^2)$$

where

$$\boldsymbol{w} = \{w_0, \ldots w_d\}$$
$$\boldsymbol{x} = \{1, x_1, \ldots x_d\}$$

Why is "1" here?



The response variable responds directly and linearly to changes in each of the inputs

Ordinary least squares regression

Given data set D

Case	X_1	X_2			X_p	Y
1	x_{11}	x_{21}			x_{p1}	y_1
2	x_{12}	x_{22}			x_{p2}	y_2
3	x_{13}	x_{23}			x_{p3}	y_3
N	x_{1N}	x_{2N}			x_{pN}	y_N

Estimation: maximizing the likelihood

$$\hat{w} = \max_w p(D|w)$$

Is equivalent to minimizing

$$RSS(w) = \sum_{i=1}^{n} (Y_i - \boldsymbol{w}^T \boldsymbol{X_i})^2$$

Matrix formulation of OLS regression

Optimality condition:

where $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = 0$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & & x_{p1} \\ 1 & x_{12} & x_{22} & & x_{p2} \\ & & & & \\ & & & & \\ 1 & x_{1N} & x_{2N} & & x_{pN} \end{pmatrix} \quad \textbf{and} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \\ \\ y_N \end{pmatrix}$$

Parameter estimates and predictions

- Least squares estimates of the parameters

$$\hat{w} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

- Predicted values

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{w} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}$$

- Linear regression belongs to the class of **linear smoothers**

Hat matrix

Why is it called so?

Degrees of freedom

Definition:

$$df(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{N} Cov(\hat{y}_i, y_i)$$

- Larger covariance → stronger connection → model can approximate data better→ model more flexible (complex)
- For linear smoothers $\hat{Y} = S(X)Y$

$$df = trace(S)$$

- For linear regression, degrees of freedom is

$$df = trace(P) = p$$

Different types of features

- Interval variables
- Numerically coded ordinal variables
  - (small=1, medium=2, large=3)
- Dummy coded qualitative variables

Basis function expansion:
If $y = w_0 + w_1 x_1 + w_2 x_1^2 + w_3 e^{-x_2} + \epsilon$,

Model becomes linear if to recompute:

$$\phi_1(x_1) = x_1$$
$$\phi_2(x_1) = x_1^2$$
$$\phi_3(x_1) = e^{-x_2}$$

Example of dummy coding:

$$x_1 = \begin{cases} 1, \text{if Jan} \\ 0, \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1, \text{if Feb} \\ 0, \text{otherwise} \end{cases}$$

.
.
.

$$x_{11} = \begin{cases} 1, \text{if Nov} \\ 0, \text{otherwise} \end{cases}$$

Basis function expansion

- In general $\phi_1(\dots)$ may be a function of several x components
- Having data given by **X,** compute new data
- $\Phi = \begin{pmatrix} 1 & \phi_1(x_{11,}\dots,x_{1p}) & .. & \phi_p(x_{11,}\dots,x_{1p}) \\ & \dots & \dots & \dots \\ 1 & \phi_1(x_{n1,}\dots,x_{np}) & \dots & \phi_p(x_{n1,}\dots,x_{np}) \end{pmatrix}$
- If doing a basis function in a model, replace **X** by Φ everywhere where **X** is used:

$$\hat{y} = \Phi(\Phi^T\Phi)^{-1}\Phi^T y$$

Linear regression in R

- fit=lm(formula, data, subset, weights,…)
  - **data** is the data frame containing the predictors and response values
  - **formula** is expression for the model
  - **subset** which observations to use (training data)?
  - **weights** should weights be used?

fit is object of class **lm** containing various regression results.

- Useful functions (many are generic, used in many other models)
  - Get details about the particular function by "?", for ex. predict.lm

```
summary(fit)
predict(fit, newdata, se.fit, interval)
coefficients(fit) # model coefficients
confint(fit, level=0.95) # CIs for model parameters
fitted(fit) # predicted values
residuals(fit) # residuals
```


An example of ordinary least squares regression

```
mydata=read.csv2("Bilexempel.csv")
fit1=lm(Price~Year, data=mydata)
summary(fit1)
fit2=lm(Price~Year+Mileage+Equipment,
data=mydata)
summary(fit2)
```

Response variable:
Requested price of used Porsche cars (1000 SEK)

```
> summary(fit1)

Call:
lm(formula = Price ~ Year, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-167683 -14683   20056   35933   72317

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -78161027    8448038  -9.252 6.00e-13 ***
Year            39246       4226   9.288 5.25e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57270 on 57 degrees of freedom
Multiple R-squared:  0.6021,    Adjusted R-squared:  0.5952
F-statistic: 86.26 on 1 and 57 DF,  p-value: 5.248e-13
```

Inputs:
X_1 = Manufacturing year
X_2 = Milage (km)
X_4 = Equipment (0 or 1)

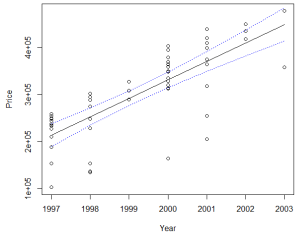
An example of ordinary least squares regression

```
> summary(fit2)

Call:
lm(formula = Price ~ Year + Mileage + Equipment, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-66223 -10525   -739  14128  65332

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.083e+07  6.309e+06  -3.302  0.00169 **
Year         1.062e+04  3.154e+03   3.366  0.00139 **
Mileage     -2.077e+00  2.022e-01 -10.269 2.14e-14 ***
Equipment    5.790e+04  1.041e+04   5.563 8.08e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29270 on 55 degrees of freedom
Multiple R-squared:  0.8997, Adjusted R-squared:  0.8942
F-statistic: 164.5 on 3 and 55 DF,  p-value: < 2.2e-16
```


An example of ordinary least squares regression

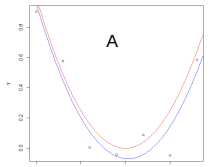
- Prediction

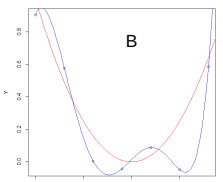
```
fitted <- predict(fit1, interval =
"confidence")

# plot the data and the fitted line
attach(mydata)
plot(Year, Price)
lines(Year, fitted[, "fit"])

# plot the confidence bands
lines(Year, fitted[, "lwr"], lty = "dotted",
col="blue")
lines(Year, fitted[, "upr"], lty = "dotted",
col="blue")
detach(mydata)
```



Ridge regression

- Problem: linear regression can overfit:
  - Take $Y := Y, X_1 = X, X_2 = X^2, \dots, X_p = X^p$ → polinomial model, fit by linear regression
  - High degree of polynomial leads to overfitting.





Ridge regression

- **Idea**: Keep all predictors but shrink coefficients to make model less complex

minimize $-loglikelihood + \lambda_0\|w\|_2^2$

→ **l_2 regularization**

  - Given that model is Gaussian, we get **Ridge regression:**

$$\hat{w}^{ridge} = \text{argmin}\left\{\sum_{i=1}^{N}(y_i - w_0 - w_1x_{1j} - \dots - w_px_{pj})^2 + \lambda\sum_{j=1}^{p}w_j^2\right\}$$

- $\lambda > 0$ is **penalty factor**

Ridge regression

Equivalent form

$$\hat{w}^{ridge} = \text{argmin}\sum_{i=1}^{N}(y_i - w_0 - w_1x_{1j} - \dots - w_px_{pj})^2$$

subject to $\sum_{j=1}^{p}w_j^2 \le s$

Solution

$$\hat{\boldsymbol{w}}^{\boldsymbol{ridge}} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{w}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}$$

Hat matrix

How do we compute degrees of freedom here?

Ridge regression

Properties

- Extreme cases:
  - $\lambda = 0$ usual linear regression (no shrinkage)
  - $\lambda = +\infty$ fitting a constant ($w = 0$ except of w_0)
- When input variables are ortogonal (not realistic), $X^TX = I \rightarrow$

$$\hat{\boldsymbol{w}}^{\text{ridge}} = \frac{1}{1+\lambda}\mathbf{w}^{\mathbf{linreg}}$$ → coefficients are equally shrunk

- **Ridge regression is particularly useful if the explanatory variables are strongly correlated to each other.**
  - Correlated variables often correspond large w→shrunk
- Degrees of freedom decrease when λ increases
  - $\lambda = 0 \rightarrow d.f. = p$

Ridge regression

Properties

- Shrinking enables estimation of regression coefficients even if the number of parameters exceeds the number of cases! ($X^TX + \lambda I$ is always nonsingular)
  - Compare with linear regression
- How to estimate λ?
  - cross-validation

Ridge regression

- Bayesian view
  - Ridge regression is just a special form of Bayesian Linear Regression with constant σ^2:

$$y \sim N(y|w_o + Xw, \sigma^2 I)$$
$$w \sim N\left(0, \frac{\sigma^2}{\lambda} I\right)$$

Theorem MAP estimate to the Bayesian Ridge is equal to solution in frequenist Ridge

$$\hat{w}^{ridge} = \left(X^T X + \lambda I\right)^{-1} X^T y$$

- In Bayesian version, we can also make inference about λ

Ridge regression

Example Computer Hardware Data Set : performance measured for various processors and also

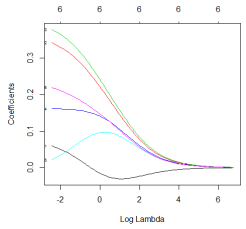
- Cycle time
- Memory
- Channels
- …

Build model predicting performance

Ridge regression

- R code: use package **glmnet** with alpha=0 (Ridge regression)
- Seeing how Ridge converges

```
data=read.csv("machine.csv", header=F)
covariates=scale(data[,3:8])
response=scale(data[, 9])

model0=glmnet(as.matrix(covariates),
response, alpha=0,family="gaussian")
plot(model0, xvar="lambda", label=TRUE)
```

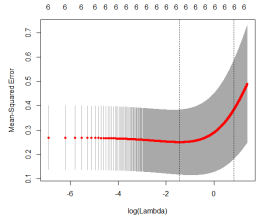


Ridge regression

- Choosing the best model by cross-validation:

```
model=cv.glmnet(as.matrix(covariates),
response, alpha=0,family="gaussian")
model$lambda.min
plot(model)
coef(model, s="lambda.min")
```

```
> coef(model, s="lambda.min")
7 x 1 sparse Matrix of class "dgC
                      1
(Intercept) -4.530442e-17
V3           3.420739e-02
V4           3.085696e-01
V5           3.403839e-01
V6           1.593470e-01
V7           5.489116e-02
V8           1.970982e-01
```



> model$lambda.min
[1] 0.046

Ridge regression

- How good is this model in prediction?

```
ind=sample(209, floor(209*0.5))
data1=scale(data[,3:9])
train=data1[ind,]
test=data1[-ind,]

covariates=train[,1:6]
response=train[, 7]
model=cv.glmnet(as.matrix(covariates), response, alpha=1,family="gaussian",
lambda=seq(0,1,0.001))
y=test[,7]
ynew=predict(model, newx=as.matrix(test[, 1:6]), type="response")

#Coefficient of determination
sum((ynew-mean(y))^2)/sum((y-mean(y))^2)

sum((ynew-y)^2)
```

Note that data are so small so numbers change much for other train/test

```
> sum((ynew-mean(y))^2)/sum((y-mean(y))^2)
[1] 0.5438148
> sum((ynew-y)^2)
[1] 18.04988
>
```


LASSO

- **Idea**: Similar idea to Ridge
- Minimize minus loglikelihood plus **linear** penalty factor→ $\mathbf{l_1}$ **regularization**
  - Given that model is Gaussian, we get **LASSO** (least absolute shrinkage and selection operator):

$$\hat{w}^{lasso} = \operatorname{argmin}\left\{\sum_{i=1}^{N}(y_i - w_0 - w_1 x_{1j} - \ldots - w_p x_{pj})^2 + \lambda \sum_{j=1}^{p}|w_i|\right\}$$

- $\lambda > 0$ is **penalty factor**

LASSO

- Equivalently

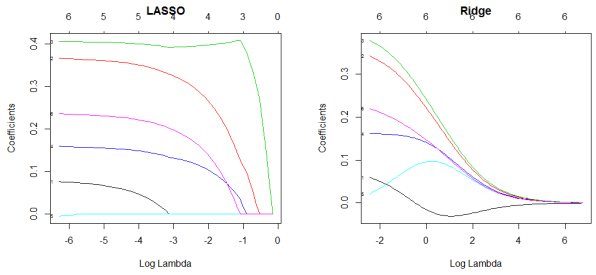
$$\hat{w}^{lasso} = \operatorname{argmin}\sum_{i=1}^{N}(y_i - w_0 - w_1 x_{1j} - \ldots - w_p x_{pj})^2$$

subject to $\sum_{j=1}^{p}|w_i| \le s$

LASSO vs Ridge

- **LASSO yields sparse solutions!**

Example Computer hardware data



LASSO vs Ridge

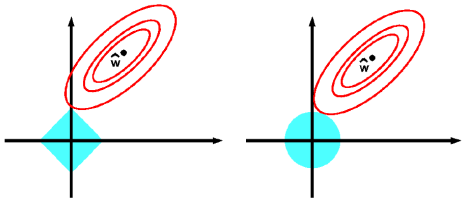
- Only 5 variables selected by LASSO

```
> coef(model, s="lambda.min")
7 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -5.091825e-17
V3           6.350488e-02
V4           3.578607e-01
V5           4.033670e-01
V6           1.541329e-01
V7           .
V8           2.287134e-01
> 
```

```
> sum((ynew-mean(y))^2)/sum((y-mean(y))^2)
[1] 0.5826904
> sum((ynew-y)^2)
[1] 16.63756
```


LASSO vs Ridge

- Why Lasso leads to sparse solutions?
  - Feasible area for Ridge is a circle (2D)
  - Feasible area for LASSO is a polygon (2D)



LASSO properies

- **Lasso is widely used when $p \gg n$**
  - Linear regression breaks down when $p > n$
  - Application: DNA sequence analysis, Text Prediction
- When inputs are orthonormal,

$$\hat{w}_i^{\text{lasso}} = sign(w_i^{linreg})\left(|w_i^{linreg}| - \frac{\lambda}{2}\right)_+$$

- No explicit formula for $\hat{w}^{lasso}$
  - Optimization algorithms used

Coding in R: use glmnet() with alpha=1

Variable selection

- .. Or “Feature selection”

Often, we do not need all features available in the data to be in the model

Reasons:

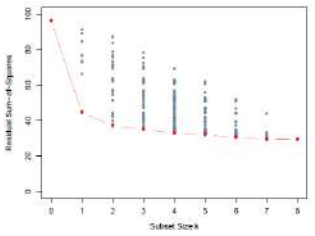
- Model can become overfitted (recall polynomial regression)
- Large number of predictors → model is difficult to use and interpret

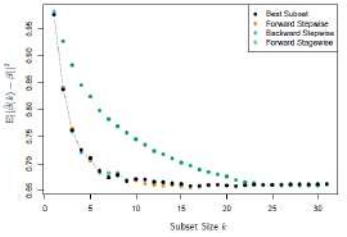
Variable selection

Alternative 1: Variable subset selection

- Best subset selection:
  - Consider different subsets of the full set of features, fit models and evaluate their quality
    - Problem: computationally difficult for p around 30 or more
    - How to choose the best model size? Some measure of predictive performance normally used (ex. AIC).
- Forward and Backward stepwise selection
  - Starts with 0 features (or full set) and then adds a feature (removes feature) that most improves the measure selected.
    - Can handle large *p* quickly
    - Does not examine all possible subsets (not the “best”)

RSS and MSE depend on k





Variable selection in R

- Use stepAIC() in MASS

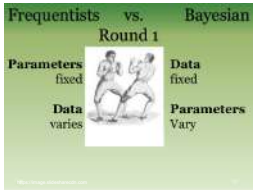
```
library(MASS)
fit <- lm(V9~.,data=data.frame(data1))
step <- stepAIC(fit, direction="both")
step$anova
summary(step)
```

```
Call:
lm(formula = V9 ~ V3 + V4 + V5 + V6 + V8, data = data.fra

Residuals:
     Min       1Q   Median       3Q      Max
-1.20232 -0.15512  0.03579  0.16567  2.42280

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.783e-17  2.574e-02   0.000   1.0000
V3           7.948e-02  2.826e-02   2.813   0.0054 **
V4           3.661e-01  4.312e-02   8.490 4.34e-15 ***
V5           4.055e-01  4.664e-02   8.695 1.18e-15 ***
V6           1.591e-01  3.394e-02   4.687 5.07e-06 ***
V8           2.360e-01  3.356e-02   7.031 3.06e-11 ***
```

```
> step <- stepAIC(fit, direction="both")
Start:  AIC=-405.35
V9 ~ V3 + V4 + V5 + V6 + V7 + V8

       Df Sum of Sq     RSS     AIC
- V7    1    0.0139 28.117 -407.25
<none>              28.103 -405.35
- V3    1    1.0819 29.185 -399.46
- V6    1    2.9385 31.041 -386.57
- V8    1    6.3150 34.418 -364.99
- V4    1    9.7492 37.852 -345.11
- V5    1   10.4837 38.586 -341.09

Step:  AIC=-407.25
V9 ~ V3 + V4 + V5 + V6 + V8

       Df Sum of Sq     RSS     AIC
<none>              28.117 -407.25
+ V7    1    0.0139 28.103 -405.35
- V3    1    1.0958 29.212 -401.26
- V6    1    3.0431 31.160 -387.77
- V8    1    6.8472 34.964 -363.70
- V4    1    9.9840 38.101 -345.74
- V5    1   10.4713 38.588 -343.08
```

Frequentist vs Bayesian

- Probabilistic Model $p(y, x, w)$
  - **Frequentists**: w is a parameter that should be estimated by model fitting
  - **Bayesians**: w is a random variable that has a prior distribution $p(w)$
    - How to set $p(w)$??



Example: Linear regression, what are parameters here?

$y \sim w_0 + \boldsymbol{wx} + e, e \sim N(0, \sigma^2)$

$y \sim N(w_0 + \boldsymbol{wx}, \sigma^2)$

An estimator

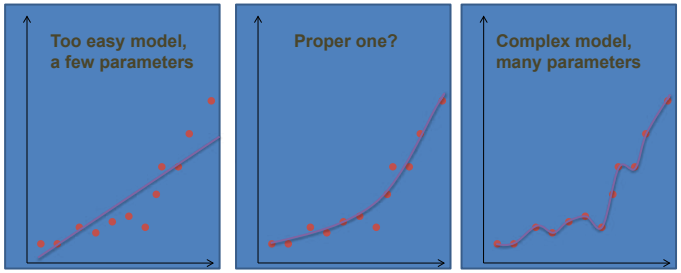
- $\widehat{\boldsymbol{w}} = \delta(D)$ (some function of your data) – an **estimator**
- Optimal parameter values?→there can be many ways to compute them (MLE, shrinkage…)
  - Compare Bayesian: given estimators $\boldsymbol{w^1}$ and $\boldsymbol{w^2}$, we **can** compare them! $p(\boldsymbol{w^1}|\boldsymbol{D}) > \boldsymbol{p}(\boldsymbol{w^2}|\boldsymbol{D})$
  - There is no easy way to compare estimators in frequentist tradition

Example: Linear regression

- Estimator 1: $\boldsymbol{w} = (X^T X)^{-1} X^T Y$ (maximum likelihood)
- Estimator 2: $\boldsymbol{w} = (0, \dots, 0, 1)$
- Which one is better?
  - A comparison strategy is needed!

Overfitting

- Complex model can overfit your data



Overfitting: solutions

- Observed: Maximum likelihood can lead to overfitting.
- Solutions
  - Selecting proper parameter values
    - Regularized risk minimization
  - Selecting proper model type, for ex. number of parameters
    - Houldout method
    - Cross-validation

Model selection

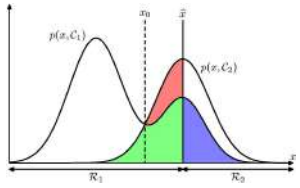
- Given a model, choose the optimal parameter values
  - Decision theory
- Define loss $L(Y, \hat{y})$
  - How much we loose in guessing true Y incorrectly
- If we know the true distribution $p(y, x|w)$ then we choose $\hat{y}$

$$\min_{\hat{y}} EL(y, \hat{y}) = \min_{\hat{y}} \int L(y, \hat{y}) p(y, x|w) dx dy$$

Model selection

Example: Spam classification

- Loss for incorrect classifying mails and spams
  - $L_{12} = 100, L_{21} = 1$



Loss functions

- How to define loss function?
  - No unique choice, often defined by application
  - **Normal practice: Choose the loss related to minus loglikelihood**

Example: Predicting the amount of the product at the storage:

$$L(Y, \hat{y}) = \begin{cases} 10 - \frac{\hat{y}}{Y}, \hat{y} \le Y \\ 1000, \hat{y} > Y \end{cases}$$

Example: Compute loss function related to

- Normal distribution

Guess why such loss function was chosen

Loss functions

- Classification problems
  - Common loss function $L(Y, \hat{y}) = \begin{cases} 0, Y = \hat{y} \\ 1, Y \neq \hat{y} \end{cases}$
  - When minimizing the loss, equivalent to misclassification rate

Model selection

- **Problem**: true model and true $\boldsymbol{w}$ are unknown→can not compute expected loss!
- How to find an optimal model?
  - Consider what expected loss (**risk**) depends on

$$R(Y, \hat{y}) = E[L(Y, \hat{y}(X, D))]$$

- Random factors:
  - D – **training set**
  - Y, X – data to be predicted **(validation set)**

Holdout method

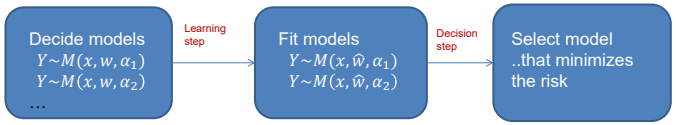
- Simplify the risk estimation:
  - Fix D as a particular training set T
  - Fix Y,X as a particular validation set V
- Risk becomes (**empirical risk**)

$$\hat{R}(y,\hat{y}) = \frac{1}{|V|}\sum_{(X,Y)\in V} L(Y,\hat{y}(X,T))$$

  - Estimator is fit by Maximum Likelihood using training set
  - Risk estimated by using validation set
  - Model with minimum empirical risk is selected

General model selection strategy

- Given data $D = \{\boldsymbol{X}_i, Y_i, i = 1 \dots n\}$



- When fitting data, Maximum Likelihood is usually used
- α_i can be different things:
  - Type of distribution
  - Number of variables in the model
  - Regulatization parameter value
  - …

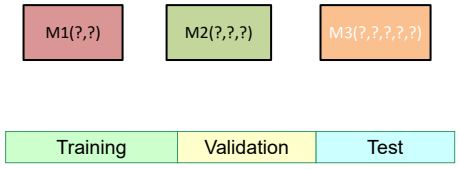
Holdout method

Divide into training, validation and test sets

- Choose proportions in some way

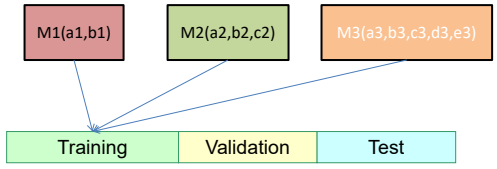
Holdout method

- Given: training, validation, test sets and models to select between



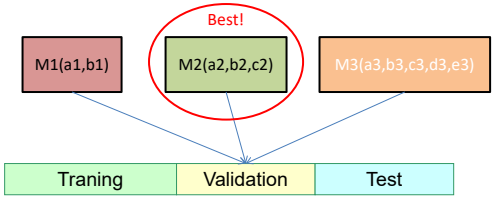
Holdout method

- Training set is to used for fitting models to the dataset by using maximum likelihood



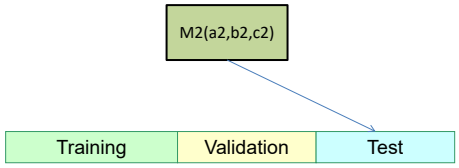
Holdout method

- Validation set is used to choose the best model (lowest risk)

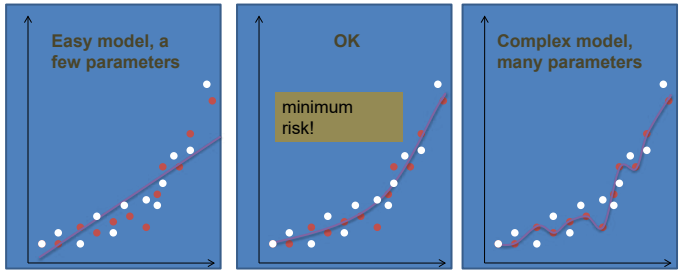


Holdout method

- Test set is used to test a performance on a new data



Holdout method



Holdout in R

- How to partition into train/test?
  - Use set.seed(12345) in the labs to get identical results

```
n=dim(data)[1]
set.seed(12345)
id=sample(1:n, floor(n*0.7))
train=data[id,]
test=data[-id,]
```

- How to partition into train/valid/test?

```
n=dim(data)[1]
set.seed(12345)
id=sample(1:n, floor(n*0.4))
train=data[id,]

id1=setdiff(1:n, id)
set.seed(12345)
id2=sample(id1, floor(n*0.3))
valid=data[id2,]

id3=setdiff(id1,id2)
test=data[id3,]
```

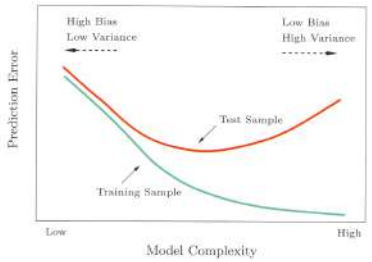
Bias-variance tradeoff

- Bias of an estimator $Bias(\hat{y}\ (x_0)) = E[\hat{y}(x_0) - f(x_0)]$, $f(x_0)$ is expected response
  - If $Bias(\hat{y}(x_0)) = 0$, the estimator is **unbiased**
  - ML estimators are asymptotically unbiased if the model is enough complex
  - However, unbiasedness does not mean a good choice!

Bias-variance tradeoff

- Assume loss is $L(Y, \hat{y}) = (Y - \hat{y})^2$

$$R(Y(x_0), \hat{y}(x_0)) = \sigma^2 + Bias^2(\hat{y}(x_0)) + Var(\hat{y}(x_0))$$



When loss is not quadratic, no such nice formula exist

Cross-validation

- Compared to holdout method:
  - Why do we use only some portion of data for training- can we use more (increase accuracy)?

Cross-validation (Estimates Err)

K-fold cross-validation (rough scheme, show picture):

1. Permute the observations randomly
2. Divide data-set in K roughly equally-sized subsets
3. Remove subset #i and fit the model using remaining data.
4. Predict the function values for subset #i using the fitted model.
5. Repeat steps 3-4 for different i
6. CV= squared difference between observed values and predicted values (another function is possible)

Cross-validation

Cross-validation

Note: if K=N then method is *leave-one-out* cross-validation.

$$\kappa : \{1, \ldots, N\} \mapsto \{1, \ldots, K\}$$

K-fold cross-validation: $CV = \frac{1}{N}\sum_{i=1}^{N} L(Y_i, \hat{y}^{-k(i)}(x_i))$

What to do if N is not a multiple of K?

Cross-validation vs Holdout

- Holdout is easy to do (a few model fits to each data)
- Cross validation is computationally demanding (many model fits)
- Holdout is applicable for large data
  - Otherwise, model selection performs poorly
- Cross validation is more suitable for smaller data

Analytical methods

- Analytical expressions to select models
  - AIC (Akaike's information criterion)

Idea: Instead of $R(Y, \hat{y}) = E[L(Y, \hat{y}(X, D))\ ]$ consider **in-sample** risk (only *Y* in *D* is random):

$$R_{in}(Y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N} E_{Y_i}\left[L\left(Y_i, \hat{y}(X, D)\right) | D, X \in D\right]$$

Analytical methods

- One can show that

$$R_{in}(Y, \hat{y}) \approx R_{train} + \frac{2}{N}\sum_{i} cov(\hat{y}_i, Y_i)$$

where $R_{train} = \frac{1}{N}\sum_{X_i Y_i \in T} L(Y_i, \hat{y}_i)$

- Recall, **degrees of freedom** $\mathrm{df(model)} = \frac{1}{\sigma^2}\sum_i cov(\hat{y}_i, Y_i)$
  - When model is linear, *df* is the number of parameters.
- If loss is defined by minus two loglikelihood,

$$AIC \equiv -2loglik(D) + 2df(model)$$

Model selection

Example Computer Hardware Data Set : performance measured for various processors and also

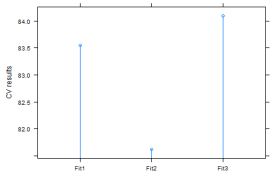
- Cycle time
- Memory
- Channels
- …

Build model predicting performance

Cross-validatation

- Try models with different predictor sets

```
data=read.csv("machine.csv", header=F)
library(cvTools)

fit1=lm(V9~V3+V4+V5+V6+V7+V8, data=data)
fit2=lm(V9~V3+V4+V5+V6+V7, data=data)
fit3=lm(V9~V3+V4+V5+V6, data=data)
f1=cvFit(fit1, y=data$V9, data=data,K=10,
foldType="consecutive")
f2=cvFit(fit2, y=data$V9, data=data,K=10,
foldType="consecutive")
f3=cvFit(fit3, y=data$V9, data=data,K=10,
foldType="consecutive")
res=cvSelect(f1,f2,f3)
plot(res)
```



Linear classification methods

Lecture 2a

Overview

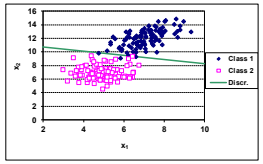
- Elements of decision theory
- Logistic regression
- Discriminant Analysis models

Classification

- Given data $D = \left((X_i, Y_i), i = 1 \dots N\right)$
  - $Y_i = Y(X_i) = C_j \in \boldsymbol{C}$
  - Class set $\boldsymbol{C} = (C_1, \dots, C_K)$

Classification problem:

- Decide $\hat{Y}(x)$ that maps **any** x into some class C_K
  - Decision boundary



Classifiers

- **Deterministic**: decide a rule that directly maps X into $\hat{Y}$
- **Probabilistic:** define a model for $P(Y = C_i|X), i = 1 \dots K$

Disanvantages of deterministic classifiers:

- Sometimes simple mapping is not enough (risk of cancer)
- Difficult to embed loss-> rerun of optimizer is often needed
- Combining several classifiers into one is more problematic
  - Algorithm A classifies as spam, Algorithm B classifies as not spam → ???
  - P(Spam|A)=0.99, P(Spam|B)=0.45 → better decision can be made

Bayesian decision theory

- Machine learning models estimate $p(y|x)$ or $p(y|x, \hat{w})$
- Transform probability into action→ which value to predict?→decision step
  - $p(Y = Spam|x) = 0.83$→do we move the mail to Junk?
  - What is more dangerous: deleting 1 non-spam mail or letting 1 spam mail enter Inbox?
- →**Loss function** or **Loss matrix**

Loss matrix

- Costs of classifying $Y = C_k$ to C_j:
  - Rows: true, columns: predicted

$$L = \|L_{ij}\|, i = 1, \dots, n, j = 1, \dots, n$$

- Example 1: 0/1-loss

$$L = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

- Example 2: Spam

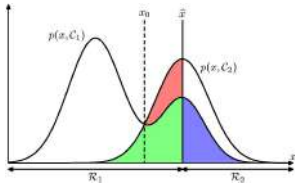
$$L = \begin{pmatrix} 0 & 100 \\ 1 & 0 \end{pmatrix}$$

Loss and decision

- Expected loss minimization
  - R_j : classify to C_j

$$EL = \sum_k \sum_j \int_{R_j} L_{kj} p(\boldsymbol{x}, C_k) d\boldsymbol{x}$$

- **Choose such R_j that EL is minimized**
- Two classes

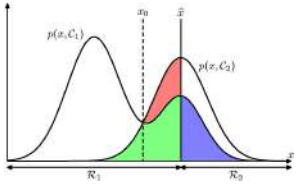


$$EL = \int_{R_1} L_{21} p(x, C_2) dx + \int_{R_2} L_{12} p(x, C_1)\, dx$$

Loss and decision

- Loss minimization

$$\min_{\hat{y}} EL(y, \hat{y}) = \min_{f} \int L(y, \hat{y}) p(y, x|w) dx dy$$



When loss is
$\begin{cases} 1, wrongly\ classified \\ 0, correctly\ classified \end{cases}$

Classify Y as
$\hat{Y} = \arg\max_c p(Y = c|X)$

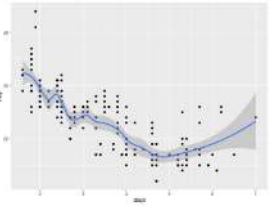
Loss and decision

- How to minimize *EL with two classes?*
- Rule:
  - $L_{12} p(x, C_1) > L_{21} p(x, C_2)$ →predict y as C_1
- 0/1 Loss: classify to the class which is more probable!

$$\frac{p(C_1|x)}{p(C_2|x)} > \frac{L_{21}}{L_{12}} \rightarrow predict\ y\ as\ C_1$$

Loss and decision

- Continuous targets: squared loss
  - Given a model $p(x,y)$, minimize
    $EL = \int L\left(y, \hat{Y}(x)\right) p(x,y)dxdy$
- Using **square loss**, the optimal is posterior mean
  $\hat{Y}(x) = \int yp(y|x)dy$



ROC curves

- Binary classification
- The choice of the thershold $\hat{x} = \frac{L_{21}}{L_{12}}$ affects prediction→ what if we don't know the loss? Which classifier is better?
- **Confusion matrix**

		PREDICTED		
		1	0	Total
TRUE	1	TP	FN	N_+
	0	FP	TN	N_-

ROC curves

- **True Positive Rates** (TPR) = **sensitivity** = **recall**
  - Probability of detection of positives: TPR=1 positives are correctly detected

$$TPR = TP/N_+$$

- **False Positive Rates** (FPR)
  - Probability of false alarm: system alarms (1) when nothing happens (true=0)

$$FPR = FP/N_-$$

- **Specificity**

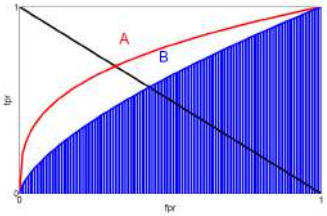
$$Specificity = 1 - FPR$$

- **Precision**

$$Precision = \frac{TP}{TP + FP}$$

ROC curves

- **ROC**=Receiver operating characteristics
- Use various thresholds, measure TPR and FPR
- Same FPR, higher TPR→ better classifier
- Best classifier = greatest Area Under Curve (**AUC**)



Types of supervised models

- **Generative models**: model $p(X|Y,w)$and $p(Y|w)$
  - Example: k-NN classification

$$p(X = x|Y = C_i, K) = \frac{K_i}{N_i V}, p(C_i|K) = \frac{N_i}{N}$$

From Bayes Theorem,

$$p(Y = C_i|x, K) = \frac{K_i}{K}$$

- **Discriminative models**: model $p(Y|X,w)$, X constant
  - Example: logistic regression
  - $p(Y = 1|\boldsymbol{w}, \boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}}$

Generative vs Discriminative

- Generative can be used to generate new data
- Generative normally easier to fit (check Logistic vs K-NN)
- Generative: each class estimated separately→ do not need to retrain when a new class added
- Discriminative models: can replace X with $\phi(X)$ (preprocessing), method will still work
  - Not generative, distribution will change
- Generative: often make too strong assumptions about $p(X|Y,w)$ → bad performance

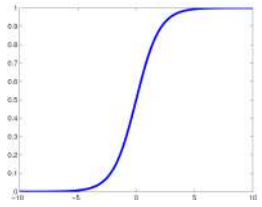
Logistic regression

- Discriminative model
- Model for binary output
  - $C = \{C_1 = 1, C_2 = 0\}$
    $p(Y = C_1|X) = sigm(\boldsymbol{w}^T\boldsymbol{x})$

$$sigm(a) = \frac{1}{1+e^{-a}}$$

- Alternatively
  $Y \sim Bernoulli(sigm(a)), a = \boldsymbol{w}^T\boldsymbol{x}$
  $sigm(a) = \frac{1}{1+e^{-a}}$

What is $P(Y = C_2|X)$?



Logistic regression

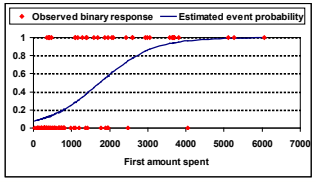
- Logistic model- yet another form

$$ln\frac{p(Y=1|X=x)}{P(Y=0|X=x)} = ln\frac{p(Y=1|X=x)}{1-P(Y=1|X=x)} = logit(p(Y=1|X=x)) = \boldsymbol{w}^T\boldsymbol{x}$$

The log of the odds is linear in *x*

- Here $logit(t) = \ln\left(\frac{t}{1-t}\right)$
- Note $p(Y|X)$ is connected to $\boldsymbol{w}^T\boldsymbol{x}$ via logit link

Example: Probability to buy more than once as function of First Amount Spend

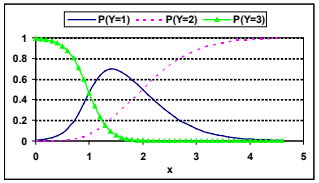


Logistic regression

- When Y is categorical,

$$p(Y = C_i|x) = \frac{e^{\boldsymbol{w}_i^T\boldsymbol{x}}}{\sum_{j=1}^{K} e^{\boldsymbol{w}_j^T\boldsymbol{x}}} = softmax(\boldsymbol{w}_i^T\boldsymbol{x})$$

- Alternatively
  $Y \sim Multinoulli\left(softmax(\boldsymbol{w}_1^T\boldsymbol{x}), \ldots softmax(\boldsymbol{w}_K^T\boldsymbol{x})\right)$



Logistic regression

Fitting logistic regression

- In binary case,

$$\log P(D|w) = \sum_{i=1}^{N} y_i \log(sigm(w^T x_i)) + (1 - y_i)\log(1 - sigm(w^T x_i))$$

  - Can not be maximized analytically, but unique maximizer exists
- To maximize loglikelihood, optimization used
  - Newton's method traditionally used (Iterative Reweighted Least Squares)
  - Steepest descent,Quasi-newton methods…

Estimation:

For new x, estimate $p(y) = [p_1, \dots p_C]$ and classify as $\arg \max_i p_i$

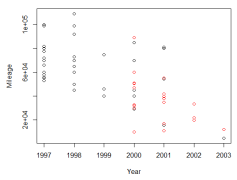
Decision boundaries of logistic regression are linear

Logistic regression

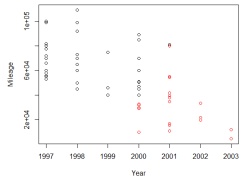
- In R, use glm() with family="binomial"
  - Predicted probabilities: predict(fit,newdata, type="response")

Example Equipment=f(Year, mileage)

Original data



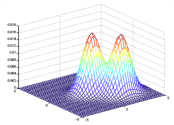
Classified data



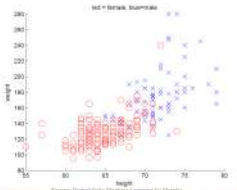
Quadratic discriminant analysis

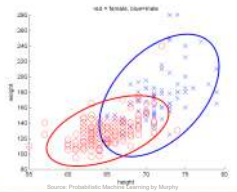
- Generative classifier
- Main assumptions:
  - $\boldsymbol{x}$ is now **random** as well as y

$$p(\boldsymbol{x}|y = C_i, \theta) = N(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$$



Unknown parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\}$

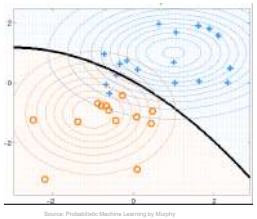




Quadratic discriminant analysis

- If parameters are estimated, classify:

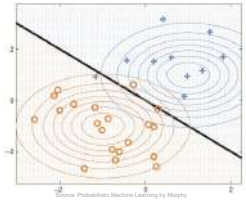
$$\hat{y}(\boldsymbol{x}) = \arg\max_c p(y = c|\boldsymbol{x}, \theta)$$



Linear discriminant analysis (LDA)

- Assumtion $\Sigma_i = \Sigma, i = 1, \dots K$
- Then $p(y = c_i|x) = softmax(w_i^T x + w_{0i})$ →exactly the same form as the logistic regression
  - $w_{0i} = -\frac{1}{2}\boldsymbol{\mu}_i^T \Sigma^{-1} \boldsymbol{\mu}_i + \log \pi_i$
  - $w_i = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i$
- Decision boundaries are linear
  - **Discriminant function**:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$



Linear discriminant analysis (LDA)

- Difference LDA vs logistic regression??
  - Coefficients will be estimated differently! (models are different)
- How to estimate coefficients
  - find MLE.

$$\hat{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c} \mathbf{x}_i, \quad \hat{\Sigma}_c = \frac{1}{N_c} \sum_{i:y_i=c} (\mathbf{x}_i - \hat{\mu}_c)(\mathbf{x}_i - \hat{\mu}_c)^T$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{c=1}^{k} N_c \hat{\Sigma}_c$$

  - Sample mean and sample covariance are MLE!
  - If class priors are parameters (**proportional priors**),

$$\hat{\pi}_c = \frac{N_c}{N}$$

LDA and QDA: code

- Syntax in R, library **MASS**

lda(formula, data, ..., subset, na.action)
- Prior – class probabiliies
- Subset – indices, if training data should be used

qda(formula, data, ..., subset, na.action)

predict(..)

LDA: output

```
resLDA=lda(Equipment~Mileage+Year, data=mydata)
 print(resLDA)
```

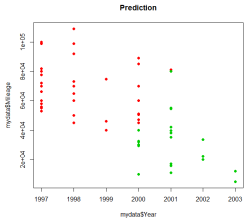
```
> print(resLDA)
Call:
lda(Equipment ~ Mileage + Year, data = mydata)

Prior probabilities of groups:
        0         1
0.6440678 0.3559322

Group means:
   Mileage     Year
0 63539.21 1998.447
1 36857.62 2000.762

Coefficients of linear discriminants:
                LD1
Mileage -1.500069e-05
Year     5.745893e-01
```


LDA: output

- Misclassified items

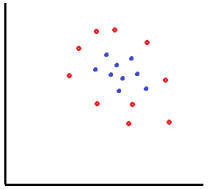
```
plot(mydata$Year, mydata$Mileage,
col=as.numeric(Pred$class)+1, pch=21,
bg=as.numeric(Pred$class)+1,
main="Prediction")
```

```
> table(Pred$class, mydata$Equipment)

    0  1
 0 31  6
 1  7 15
```



LDA versus Logistic regression

- Generative classifiers are easier to fit, discriminative involve numeric optimization
- LDA and Logistic have same model form but are fit differently
- LDA has stronger assumptions than Logistic, some other generative classifiers lead also to logistic expression
- New class in the data?
  - Logistic: fit model again
  - LDA: estimate new parameters from the new data
- Logistic and LDA: complex data fits badly unless interactions are included



LDA versus Logistic regression

- LDA (and other generative classifiers) handle missing data easier
- Standardization and generated inputs:
  - Not a problem for Logistic
  - May affect the performance of the LDA in a complex way
- Outliers affect Σ → LDA is not robust to gross outliers
- LDA is often a good classification method even if the assumption of normality and common covariance matrix are not satisfied.

Naïve Bayes classifiers
Decision trees

Lecture 2b

Naive Bayes classifiers: motivation

- Consider n labeled text documents
  - $Y = \{0,1\},\ 0 =$ "$Science\ fiction$", $1 =$ "$Comedy$"
  - $X = \{X_1, \dots X_{100}\}$ does the document contain the keyword (0=No,1=Yes)
    - X_1 corr. "space", X_2 corr. "fun",...

- Want to classify a new document

Naive Bayes classifiers: motivation

Idea: use Bayes classifier

$$p(Y = y|X) = \frac{P(X|Y = y)P(Y = y)}{\sum_j P(X|Y = y_j)P(Y = y_j)}$$

Chance of observing a given combination of words in science fiction

Proportion of science fiction documents

Naive Bayes classifiers: motivation

- Attempt 1:
  - Model $P(X = (x_1, \dots x_p)|Y = y_i)$ and $P(Y = y_i)$ as unknown parameters
  - Use data to derive those with Maximum Likelihood
  - Classify by use of the posterior distribution
- How many parameters?
  - How many different combinations of X? 2^p
  - Amount of $P(X = (x_1, \dots x_p)|Y = y_i)$ is $2 * 2^p - 2$
    - Probabilities for each Y sum up to one
- If $p = 100$, 10^{30} parameters need to be estimated→ ouch!

Naive Bayes classifiers

- Naive Bayes assumption: conditional independence

$$P(X = (x_1, \dots x_p)|Y = y) = \prod_{i=1}^{p} P(X_i = x_i|Y = y)$$

- How many parameters now?
  - $P(X_i = x_i|Y = y), i = 1, \dots p, x_i = \{0,1\}, y = \{0,1\}$ $2 * p$
- Is Naive Bayes assumption always valid?
  - P(Space,ship|SciFi)= P(Space|SciFi)*P(Ship|SciFi) ?

Naive Bayes classifiers - discrete inputs

- Given $D = \{(X_{m1}, \dots X_{mp}, Y_m),\ m = 1, \dots n\}$
- Assume $X_i \in \{x_1, \dots x_j\}, i = 1, \dots p, Y \in \{y_1 \dots y_K\}$
- Denote $\theta_{ijk} = p(X_i = x_j|Y = y_k)$
  - How many parameters? $(J-1)Kp$
- Denote $\pi_k = p(Y = y_k)$
- Maximum likelihood: assume θ_{ijk} and π_k are constants
  - $\hat{\theta}_{ijk} = \frac{\#\{X_i = x_j \,\&\, Y = y_k\}}{\#\{Y = y_k\}}$
  - $\hat{\pi}_k = \frac{\#\{Y = y_k\}}{n}$
  - Classification using 0-1 loss: $\hat{Y} = \arg\max_y p(Y = y|X)$

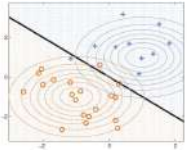
Naive Bayes classifiers - discrete inputs

- **Example** Loan decision
  - Classify a person: Home Owner=No, Single=Yes

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naive Bayes – continuous inputs

- X_i are continuous
- Assumption A: $x_j|y = C$ are univariate Gaussian
  - $p(x_j|y = C_i, \theta) = N(x_j|\mu_{ij}, \sigma_{ij}^2)$
- Therefore $p(\boldsymbol{x}|y = C_i, \theta) = N(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$
  - $\boldsymbol{\Sigma_i}$=diag$(\sigma_{i1}^2, \dots, \sigma_{ip}^2)$

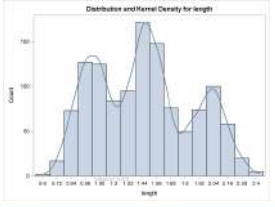


- Naive bayes is a special case of LDA (given A)
  - →MLE are means and variances (per class)

Naive Bayes – continuous inputs

- Assumption B: $p(x_j|y = C)$ are unknown functions of x_j that can be estimated from data
  - Nonparametric density estimation (kernel for ex.)

1. Estimate $p(X_i = x_j|Y = y_k)$ using nonparametric methods
2. Estimate $p(Y = y_k)$ as class proportions
3. Use Bayes rule and 0-1 loss to classify



Naive Bayes in R

- naiveBayes in package **e1071**

Example: Satisfaction of householders with their present housing circumstances

```
library(MASS)
library(e1071)
n=dim(housing)[1]
ind=rep(1:n, housing[,5])
housing1=housing[ind,-5]

fit=naiveBayes(Sat~., data=housing1)
fit

Yfit=predict(fit, newdata=housing1)
table(Yfit,housing1$Sat)
```

```
> table(Yfit,housing1$Sat)

Yfit      Low Medium High
  Low     294    162  144
  Medium  20     23   20
  High    253    261  504
```

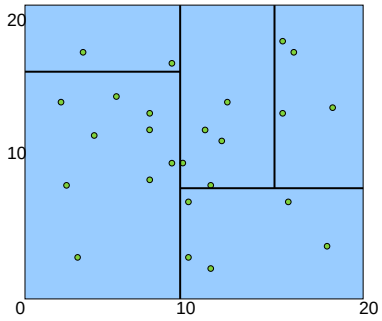

Decision trees

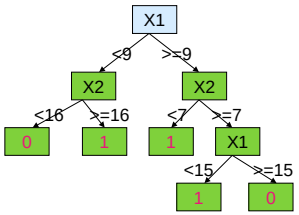
Idea

Split the domain of feature set into the set of hypercubes (rectangles, cubes) and define the target value to be constant within each hypercube

- Regression trees:
  - Target is a continuous variable
- Classification trees
  - Target is a class (qualitative) variable

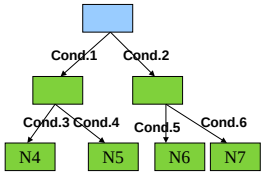
Classification tree toy example



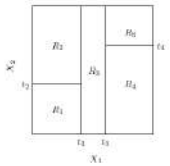


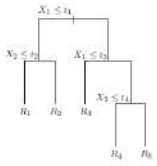
Definitions

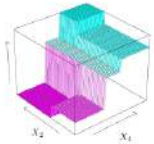
- Root node
- Nodes
- Leaves (terminal nodes)
- Parent node, child node
- Decision rules
- A value is assigned to the leaves



Regression tree toy example







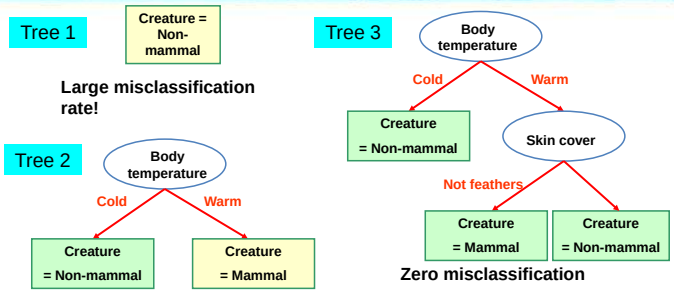
A classification problem

Create a classification tree that would describe the following patterns

ID	x1	x2	x3	x4	x5	x6	x7	y

Name	Body temperature	Skin cover	Gives birth	Aquatic creature	Aerial creature	Has legs	Hibernates	Class label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	non-mammal
salmon	cold-blooded	scales	no	yes	no	no	no	non-mammal
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	non-mammal
komodo	cold-blooded	scales	no	no	no	yes	no	non-mammal
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	non-mammal
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
shark	cold-blooded	scales	yes	yes	no	no	no	non-mammal
turtle	cold-blooded	scales	no	semi	no	yes	no	non-mammal
penguin	warm-blooded	feathers	no	semi	no	yes	no	non-mammal
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	non-mammal
salamander	cold-blooded	none	no	semi	no	yes	yes	non-mammal

Several solutions



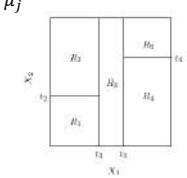
Green boxes represent pure nodes =nodes where observed values are the same

Decision trees

- A tree $T =< r_i, s_{r_i}, R_j, i = 1 \dots S, j = 1 \dots L >$
  - $x_{r_i} \leq s_{r_i}$ splitting rules (conditions), S- their amount
  - R_j-terminal nodes, L- their amount
  - labels μ_j in each terminal node

Model:

- $Y|T$ for R_j comes from exponential family with mean μ_j
- Fitting by MLE:
  - Step 1: Finding optimal tree
  - Step 2: Finding optimal labels in terminal nodes



Decision trees

Example:

- **Normal model** leads to regression trees
  - Objective: MSE
- **Multinoulli model** leads to classification trees
  - Objective: cross-entropy (deviance)

Classification trees

- Target is categorical
- Classification probability $p_{mk} = p(Y = k|X \in R_m)$ is estimated for every class in a node
- How to estimate p_{mk} for class k and node R_m?

Class proportions

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

- For any node (leave), a label can be assigned

$$k(m) = \arg\max_k \hat{p}_{mk}$$

Classification trees

- Impurity measure $Q(R_m)$
  - R_m is a tree node (region)
  - Node can be split unless it is pure

Misclassification error: $\frac{1}{N_m}\sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$

Gini index: $\sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$

Cross-entropy or deviance: $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$.

- Note: In many sources, deviance is $Q(R_m)$ $N(R_m)$

Example: Cross –entropy is MLE of $Y_j|T \sim Multinomial(p_{j1}, \ldots p_{jc})$

Fitting regression trees: CART

Step 1: Finding optimal tree: grow the tree in order to minimize global objective

1. Let C_0 be a hypercube containing all observations
2. Let queue C={C_0}
3. Pick up some C_j from C and find a variable X_j and value s that split C_j into two hypercubes

$$R_1(j,s) = \{X|X_j \leq s\} \text{ and } R_2(j,s) = \{X|X_j > s\}$$

and solve

$$\min_{j,s}[N_1 Q(R_1) + N_2 Q(R_2)]$$

4. Remove C_j from C and add R_1 and R_2
5. Repeat 3-4 as many times as needed (or until each cube has only 1 observation)

CART: comments

- Greedy algorithm (optimal tree is not found)
- The largest tree will interpolate the data → large trees = overfitting the data
- Too small trees=underfitting (important structure may not be captured)
- Optimal tree length?

Optimal trees

- **Postpruning**

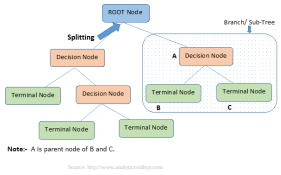
Weakest link pruning:

1. Merge two leaves that have smallest *N(parent)*Q(parent)-N(leave1)Q(leave1)-N(leave2)Q(leave2)*
2. For the current tree T, compute

$$I(T) = \sum_{R_i \in leaves} N(R_i)Q(R_i) + \alpha|T|$$

$|T|$ =#leaves

3. Repeat 1-2 until the tree with one leave is obtained
4. Select the tree with smallest I(T)



How to find the optimal α? Cross validation!

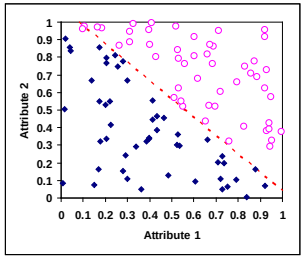
Decision trees: comments

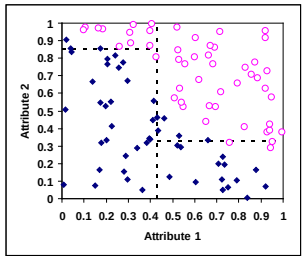
- Similar algorithms work for regression trees – replace $N \cdot Q(R)$ by $SSE(R)$
- Easy to interpret
- Easy to handle all types of features in one model
- **Automatic variable selection**
- Relatively robust to outliers
- Handle large datasets

- Trees have high variance: a small change in response→ totally different tree
- Greedy algorithms → fit may be not so good
- Lack of smoothness

Decision trees: issues

- Large trees may be needed to model an easy system:





Decision trees in R

- **tree** package
  - Alternative: **rpart**

tree(formula, data, weights, control, split = c("deviance", "gini"), …)

print(), summary(), plot(), text()

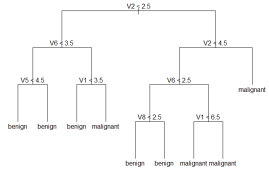
Example: breast cancer as a function av biological measurements

```
library(tree)
n=dim(biopsy)[1]
fit=tree(class~., data=biopsy)
plot(fit)
text(fit, pretty=0)
fit
summary(fit)
```


Decision trees in R

- Adjust the splitting in the tree with *control* parameter (leaf size for ex)

```
> fit
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 683 884.400 benign ( 0.650073 0.349927 )
   2) V2 < 2.5 418 108.900 benign ( 0.971292 0.028708 )
     4) V6 < 3.5 395  25.130 benign ( 0.994937 0.005063 )
       8) V5 < 4.5 389   0.000 benign ( 1.000000 0.000000 ) *
       9) V5 > 4.5 6   7.638 benign ( 0.666667 0.333333 ) *
     5) V6 > 3.5 23  31.490 benign ( 0.565217 0.434783 )
      10) V1 < 3.5 11   0.000 benign ( 1.000000 0.000000 ) *
      11) V1 > 3.5 12  10.810 malignant ( 0.166667 0.833333 ) *
   3) V2 > 2.5 265 217.900 malignant ( 0.143396 0.856604 )
     6) V2 < 4.5 90 120.300 malignant ( 0.388889 0.611111 )
      12) V6 < 2.5 30  27.030 benign ( 0.833333 0.166667 )
        24) V8 < 2.5 19   0.000 benign ( 1.000000 0.000000 ) *
        25) V8 > 2.5 11  15.160 benign ( 0.545455 0.454545 ) *
      13) V6 > 2.5 60  54.070 malignant ( 0.166667 0.833333 )
        26) V1 < 6.5 28  35.160 malignant ( 0.321429 0.678571 ) *
        27) V1 > 6.5 32   8.900 malignant ( 0.031250 0.968750 ) *
     7) V2 > 4.5 175  30.350 malignant ( 0.017143 0.982857 ) *
```




```
> summary(fit)

Classification tree:
tree(formula = class ~ ., data = biopsy)
Variables actually used in tree construction:
[1] "V2" "V6" "V5" "V1" "V8"
Number of terminal nodes:  9
Residual mean deviance:  0.1603 = 108 / 674
Misclassification error rate: 0.03221 = 22 / 683
```

Decision trees in R

- Misclassification results

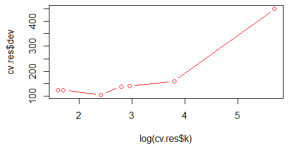
```
Yfit=predict(fit, newdata=biopsy, type="class")
table(biopsy$class,Yfit)
```

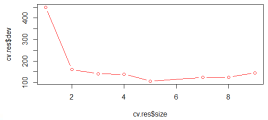
```
> table(biopsy$class,Yfit)
           Yfit
            benign malignant
  benign       440        18
  malignant      7       234
```


Decision trees in R

- Selecting optimal tree by penalizing
  - Cv.tree()

```
set.seed(12345)
ind=sample(1:n, floor(0.5*n))
train=biopsy[ind,]
valid=biopsy[-ind,]

fit=tree(class~., data=train)
set.seed(12345)
cv.res=cv.tree(fit)
plot(cv.res$size, cv.res$dev, type="b",
col="red")
plot(log(cv.res$k), cv.res$dev,
type="b", col="red")
```



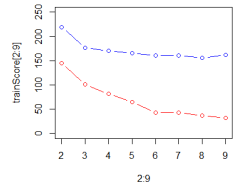


What is optimal number of leaves?

Decision trees in R

- Selecting optimal tree by train/validation

```
fit=tree(class~., data=train)

trainScore=rep(0,9)
testScore=rep(0,9)

for(i in 2:9) {
  prunedTree=prune.tree(fit,best=i)
  pred=predict(prunedTree, newdata=valid,
type="tree")
  trainScore[i]=deviance(prunedTree)
  testScore[i]=deviance(pred)
}
plot(2:9, trainScore[2:9], type="b", col="red",
ylim=c(0,250))
points(2:9, testScore[2:9], type="b", col="blue")
```



What is optimal number of leaves?

Decision trees in R

- Final tree: 5 leaves

```
finalTree=prune.tree(fit, best=5)
Yfit=predict(finalTree, newdata=valid,
type="class")
table(valid$class,Yfit)
```

```
> table(valid$class,Yfit)
           Yfit
            benign malignant
  benign       222         8
  malignant      6       114
```

Generalized Linear Models. Uncertainty estimation

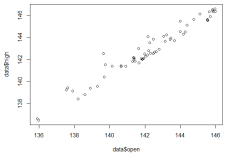
Lecture 2c

Moving beyond typical distributions

- We know how to model
  - Normally distributed targets -> linear regression
  - Bernoulli and Multinomial targets→logistic regression
  - What if target distribution is more complex?

Example 1: Daily Stock prices NASDAQ
- Open
- High (within day)

Does it seem that the error is normal here?



Example 2: Number of calls to bank
- Y=Number of calls
- X= time

Endless amount of classes → multinomial does not work… (Poisson)

Exponential family

- More advanced error distributions are sometimes needed!
- Many distributions belong to **exponential** family:
  - Normal, Exponential, Gamma, Beta, Chi-squared..
  - Bernoulli, Multinoulli, Poisson…

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})e^{(\boldsymbol{\eta}^T u(\boldsymbol{x}))}$$

- Easy to find MLE and MAP
- Non-exponential family distributions: uniform, Student t

Example: Bernoulli

Generalized linear models

- Assume Y from the exponential family
- **Model** is $Y \sim EF(\mu, \dots),\ f(\mu) = \boldsymbol{w}^T\boldsymbol{x}$
  - Alt $\mu = f^{-1}(\boldsymbol{w}^T\boldsymbol{x})$
  - f^{-1} is activation function
  - f is link function (in principle, arbitrary)
- Arbitrary f will lead to (s – dispersion parameter)

$$p(y|w,s) = h(y,s)g(\boldsymbol{w},\boldsymbol{x})e^{\frac{b(\boldsymbol{w},\boldsymbol{x})y}{s}}$$

- If f is a canonical link, then

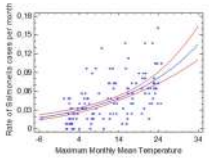
$$p(y|w,s) = h(y,s)g(\boldsymbol{w},\boldsymbol{x})e^{\frac{(\boldsymbol{w}^T\boldsymbol{x})y}{s}}$$

Generalized linear models

- Canonical links are normally used
  - MLE computations simplify
  - MLE $\hat{w} = F(X^TY)$ → computations do not depend on all data but rather a summary (sufficient statistics)→ computations speed up

Example: Poisson regression

$f^{-1}(\mu) = e^{\mu}, Y \sim Poisson\left(e^{\boldsymbol{w}^T\boldsymbol{x}}\right)$

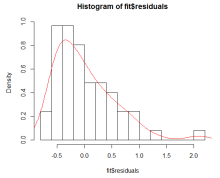


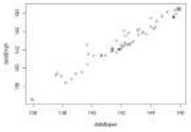
Generalized linear model: software

- Use **glm**(formula, family, data) in R

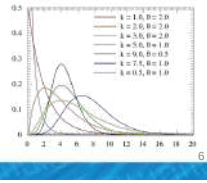
Example: Daily Stock prices NASDAQ
- Open
- High (within day)

1. Try to fit usual linear regression, study histogram of residuals



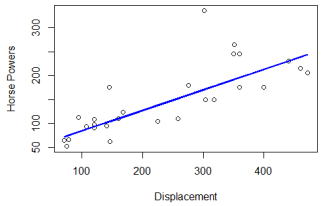


Gamma distribution: Wikipedia



Least absolute deviation regression

- Model $Y \sim Laplace(w^TX, b)$
  - Member of exponential family
- Equivalent to minimizing sum of absolute deviations
- Properties
  - Robust to outliers
  - Sensitive to changes in data
  - Multiple solutions possible



- R: package **L1pack**

Probabilistic models

- Why it is beneficial to assume a **probabilistic** model?
- A common approach to modelling in CS and engineering:

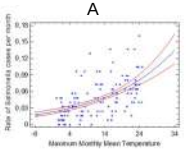
$$y = f(x, w)$$

- f is known, w is unknown
- Fit model to data with least squares, optimization or ad hoc→ find w

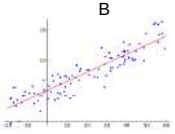
Probabilistic models

Arguments against deterministic models:
- The model does not really describe actual data (error is not explained)
  - No difference between modelling data A (Poisson) and B (Normal)
  - Estimation strategy for A is not good for B
- The model typically gives a **deterministic answer**, no information about uncertainty
  - "…The exchange rate tomorrow will be 8.22 …" 😨

A



B



Probabilistic models

Probabilistic model

$$Y \sim Distribution(f(x,w),\theta)$$

- Data is fully explained (error as well)
- Automatic principle for finding parameters: MLE , MAP or Bayes theorem
- Automatic principle for finding uncertainty (conf. limits)
  - **Bootstrap**
  - Posterior probability
- Possibility to generate new data of the same type
  - Further testing of the model

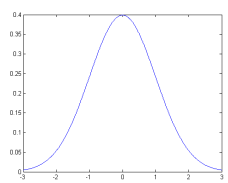
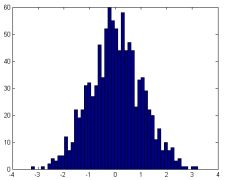
Uncertainty estimation

- Given estimator $\hat{f} = \hat{f}(x, D)$ (or $\hat{\boldsymbol{\alpha}} = \delta(D)$), how to estimate the uncertainty?

- Answer 1: if the distribution for data *D* is given, compute analytically the distribution for the estimator→ derive confidence limits
  - Often difficult
  - Example: In simple linear regression, $\hat{\alpha}$ follows *t* distribution

- Answer 2: Use **bootstrap**

The bootstrap: general principle

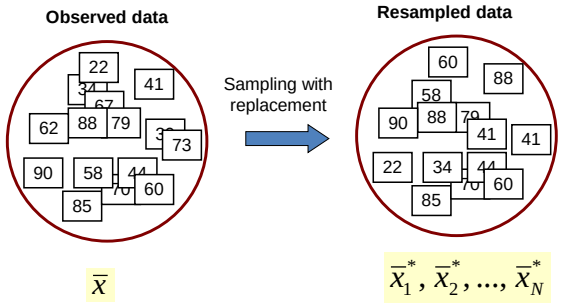



We want to determine uncertainty of $\hat{f}(D, X)$

1. Generate many different D_i from their distribution
2. Use histogram of $\hat{f}(D_i, X)$ to determine confidence limits→ unfortunately can not be done (distr of *D is often unknown)*

Instead: Generate many different D_i^* from the empirical distribution (histogram)

Nonparametric bootstrap



$\bar{x}$ $\quad \bar{x}_1^*, \bar{x}_2^*, \ldots, \bar{x}_N^*$

Nonparametric bootstrap

Given estimator $\hat{w} = \hat{f}(D)$

Assume $X \sim F(X, w)$, F and w are unknown

1. Estimate $\hat{w}$ from data **D**=(X_1,…X_n)
2. Generate $\mathbf{D_1}$ =(X^*_1,…X^*_n) by sampling with replacement
3. Repeat step 2 B times
4. The distribution of w is given by $\hat{f}(D_1)$, … $\hat{f}(D_B)$

Nonparametric bootstrap can be applied to any deterministic estimator, distribution-free

Parametric bootstrap

Given estimator $\hat{w} = \hat{f}(D)$

Assume $X \sim F(X, w)$, F is known and w is unknown

1. Estimate $\hat{w}$ from data **D**=(X_1,…X_n)
2. Generate $\mathbf{D_1}$ =(X^*_1,…X^*_n) by generating from $F(X, \hat{w})$
3. Repeat step 2 B times
4. The distribution of w is given by $\hat{f}(D_1)$, … $\hat{f}(D_B)$

Parametric bootstrap is **more** precise if the distribution form is correct

Uncertainty estimation

1. Get D_1 , … D_B by bootstrap
2. Use $\hat{f}(D_1)$, … $\hat{f}(D_B)$ to estimate the uncertainty
   - Boostrap percentile
   - Bootstrap Bca
   - …

- Bootstrap works for all distribution types
- Can be bad accuracy for small data sets $n < 40$ (empirical is far from true)
- Parametric bootstrap works even for small samples

Bootstrap confidence intervals

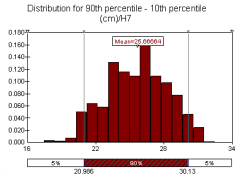
- To estimate 100(1-α) confidence interval for w

Bootstrap percentile method

1. Using bootstrap, compute $\hat{f}(D_1)$, … $\hat{f}(D_B)$, sort in ascending order, get w_1 … w_B
2. Define A_1=ceil(B α/2), A_2=floor(B-B α/2)
3. Confidence interval is given by

$$\left(w_{A_1}, w_{A_2}\right)$$

Look at the plot…



Bootstrap: regression context

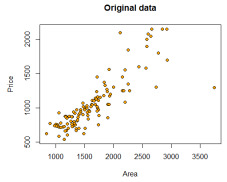
- Model $Y \sim F(X, w)$
- Data $\mathrm{D} = \{(Y_i, X_i), i = 1, \ldots, n\}$
- Idea: produce several bootstrap sets that are similar to D

Nonparametric bootstrap:

1. Using observation set **D**, sample **pairs** (X_i, Y_i) with replacement and get bootstrap sample $\mathbf{D_1}$
2. Repeat step 1 B times→ get $\mathbf{D_{1,\ldots}}$ $\mathbf{D_B}$

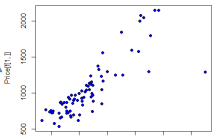
Uncertainty estimation

Example: Albuquerque dataset:
Y=Price of House
X=Area (sqft)

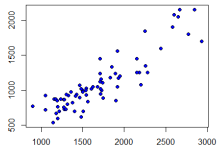


We sample data index, from {1…N}

D_1



D_{50}



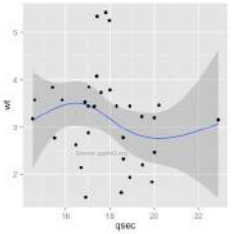
Bootstrap: regression context

Parametric bootstrap

1. Fit a model to D → get $\hat{w}(D)$.
2. Set $X_i^* = X_i$, generate $Y_i^* \sim F(X_i, \hat{w})$.
3. $D_i = \{(X_i^*, Y_i^*), i = 1, \dots, n\}$
4. Repeat step 2 B times

Confindence intervals in regression

- Given $Y \sim Distribution(y|x,w), EY|X = \mu|x = f(x,w)$
  - Example: $Y \sim N(w^T x, \sigma^2), \mu|x = f(x,w) = w^T x$
- Estimate intervals for $\mu|x = f(x,w)$ for many X, combine in a **confidence band**
- What is estimator?
  - $\mu|x = f(x,w)$



Confindence intervals in regression

Estimation

1. Compute $D_1, \dots D_B$ using a bootstrap
2. Fit model to $D_1, \dots D_B$ →estimate $\hat{W}_1, \dots \hat{W}_B$
3. For a given X, compute $f(X, \hat{w}_1), \dots f(X, \hat{w}_B)$ and estimate confidence interval by (percentile method)
4. Combine confidence intervals in a band

Bootstrap: R

- Package **boot**
  - **Functions:**
    - boot()
    - boot.ci() – 1 parameter
    - envelope() – many parameters
- Random random generation for parametic bootstrap:
  - Rnorm()
  - Runif()
  - …

boot(data, statistic, R, sim = "ordinary", ran.gen = function(d, p) d, mle = NULL,…)

Bootstrap: R

Nonparametric bootstrap:

- Write a function *statistic* that depends on *dataframe* and *index* and returns the estimator

```
library(boot)
data2=data[order(data$Area),]#reordering data according to Area

# computing bootstrap samples
f=function(data, ind){
  data1=data[ind,]# extract bootstrap sample
  res=lm(Price~Area, data=data1) #fit linear model
  #predict values for all Area values from the original data
  priceP=predict(res,newdata=data2)
  return(priceP)
}
res=boot(data2, f, R=1000) #make bootstrap
```


Bootstrap: R

Parametric bootstrap:

- Compute value *mle* that estimates model parameters from the data
- Write function *ran.gen* that depends on *data* and *mle* and which generates new data
- Write function *statistic* that depend on *data* which will be generated by *ran.gen* and should return the estimator

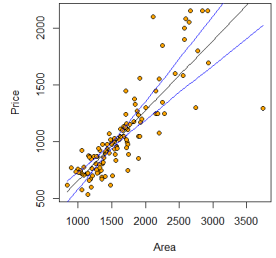
Bootstrap

```
mle=lm(Price~Area, data=data2)

rng=function(data, mle) {
  data1=data.frame(Price=data$Price, Area=data$Area)
  n=length(data$Price)
#generate new Price
  data1$Price=rnorm(n,predict(mle, newdata=data1),sd(mle$residuals))
  return(data1)
}

f1=function(data1){
  res=lm(Price~Area, data=data1) #fit linear model
  #predict values for all Area values from the original data
  priceP=predict(res,newdata=data2)
  return(priceP)
}

res=boot(data2, statistic=f1, R=1000, mle=mle,ran.gen=rng, sim="parametric")
```


Uncertainty estimation: R

- Bootstrap cofidence bands for linear model

```
e=envelope(res) #compute confidence bands

fit=lm(Price~Area, data=data2)
priceP=predict(fit)

plot(Area, Price, pch=21, bg="orange")
points(data2$Area,priceP,type="l") #plot fitted line

#plot cofidence bands
points(data2$Area,e$point[2,], type="l", col="blue")
points(data2$Area,e$point[1,], type="l", col="blue")
```



Prediction bands

- Confidence interval for Y|X= interval for mean $EY|X$
- Prediction interval for Y|X= interval for $Y|X$

$$Y \sim Distribution(x, w)$$

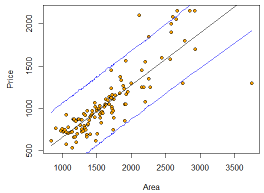
Prediction band for parametric bootstrap

1. Run parametric bootstrap and get $D_1, \dots D_B$
2. Fit the model to the data and get $\hat{w}(D_1), \dots \hat{w}(D_B)$
3. For each X, generate from $Distribution\left(X, \hat{w}(D_1)\right), \dots$ $Distribution\left(X, \hat{w}(D_B)\right)$ and apply percentile method
4. Connect the intervals→get the band

Estimation of the model quality

Example: parametric bootstrap

```
mle=lm(Price~Area, data=data2)

f1=function(data1){
  res=lm(Price~Area, data=data1) #fit
linear model
  #predict values for all Area values
from the original data
  priceP=predict(res,newdata=data2)
  n=length(data2$Price)
  predictedP=rnorm(n,priceP,
sd(mle$residuals))
  return(predictedP)
}
res=boot(data2, statistic=f1, R=10000,
mle=mle,ran.gen=rng, sim="parametric")
```



Why wider band?

Lecture 2d

Latent variable models

Overview

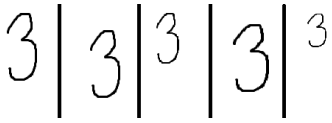
- Principal Component Analysis (PCA)
- Probabilistic PCA
- Independent component analysis (ICA)

Latent variables

- Sometimes data depends on the variables we can not measure (hard to measure)
  - Answers on the test depend on Intelligence
  - Brain activity in the brain is measured by sensors
  - Stock prices depend on market confidence

Latent variables

- Latent factor discovered→ data storage may decrease a lot
- Latent factors
  - Center
  - Scaling
- Original vs compressed
  - 100x100x5=50000
  - 100x100+2*5+2*5=10020



Principal Component Analysis (PCA)

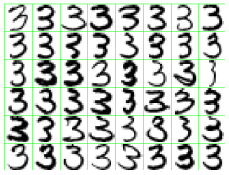
- *PCA* is a technique for reducing the complexity of high dimensional data
- It can be used to approximate high dimensional data with a few dimensions (latent features) –> much less data to store
- New variables might have a special interpretation

Applications

- Image recognition
- Information compression
- Subspace clustering
- …

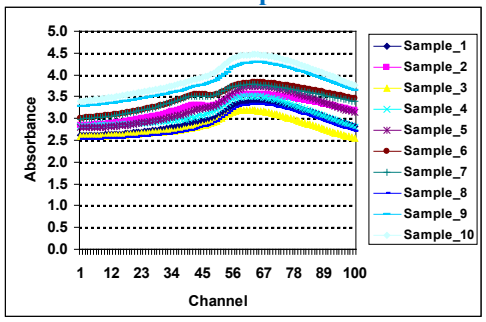
Principal Component Analysis (PCA)

- Example 1: Hadwritten digits
  - Can we get a more compact summary?



Absorbance records for ten samples of chopped meat

Parallel coordinate plot for "FAT"



1 target (fat)

100 features (absorbance at 100 wavelengths or channels)

The features are strongly correlated to each other

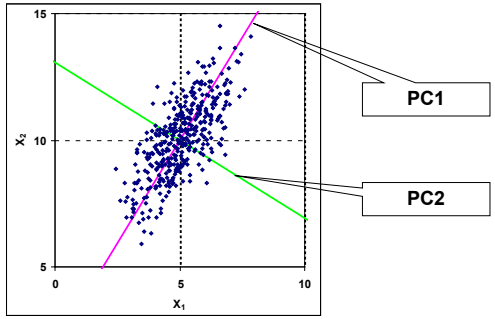
Principal components analysis

Idea: Introduce a new coordinate system (PC1, PC2, …) where

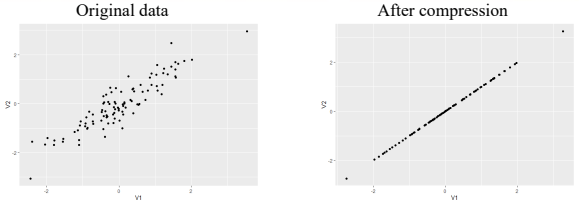
- The first principal component (PC1) is the direction that maximizes the variance of the projected data
- The second principal component (PC2) is the direction that maximizes the variance of the projected data after the variation along PC1 has been removed
- The third principal component (PC3) is the direction that maximizes the variance of the projected data after the variation along PC1 and PC2 has been removed
- ….

In the new coordinate system, coordinates corresponding to the last principal components are very small → can take away these columns

Principal Component Analysis - two inputs



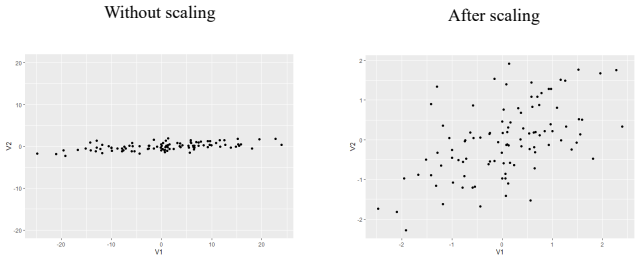
PCA- after reducing dimensionality



- Data became approximate (but less data to store)
- $PC_1, \dots PC_M$ are actually eigenvections of **sample covariance** (first largest eigenvalue,…,Mth largest egenvalue)

PCA and scaling

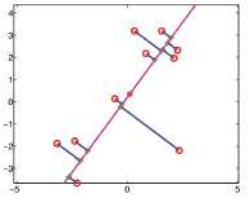
- Do we need to scale features?



PCA: another view

- Aim: minimize the distance between the original and projected data

$$\min_{U_M} \sum_{i=1}^{N} \|x_n - \tilde{x}_n\|^2$$



PCA: computations

Data $D = \left\| \mathbf{x}_1 \ \mathbf{x}_2 \ \dots \mathbf{x}_p \right\|, \quad \mathbf{x}_i = (x_{i1}, \dots, x_{in})$

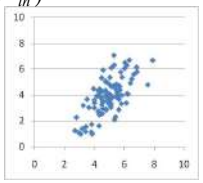
1. Centred data

$$X = \left\| \mathbf{x}_1 - \overline{\mathbf{x}}_1 \ \mathbf{x}_2 - \overline{\mathbf{x}}_2 \ \dots \mathbf{x}_p - \overline{\mathbf{x}}_p \right\|,$$

2. Covariance matrix

$$\boldsymbol{S} = \frac{1}{N} \boldsymbol{X}^T \boldsymbol{X}$$

3. Search for eigenvectors and eigenvalues of **S**



	Column 1	Column 2
Column 1	0.951	0.905
Column 2	0.905	1.883

PCA: computations

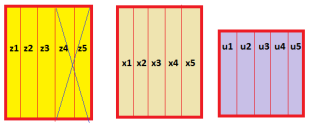
4. Coordinates of any data point x=(x_1…x_p) in the new coordinate system:
$z = (z_1, \dots z_n),\ z_i = x^T u_i$

Matrix form: $Z = X\ U$

5. Discard principle components after some *M:*
$Z = X\ U_M$

6. New data will have dimensions N x M instead of N x p

Getting approximate original data:
$\tilde{X} = Z U_M^T$

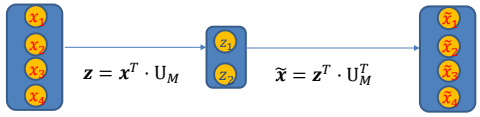


Store: N x M+ p x M instead N x p

100*50 vs
100*4+50*4

PCA: computations

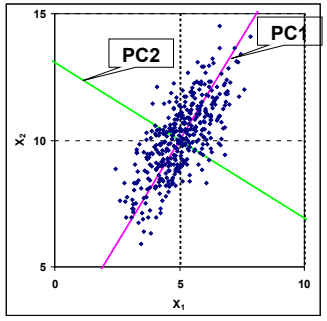
- PCA makes a **linear** compression of features



R^p $\quad R^M, M < p \quad$ R^p

$$\min_{U_M} \sum_{i=1}^{N} \|x_n - \tilde{x}_n\|^2$$

Principal Component Analysis



Eigenanalysis of the Covariance Matrix

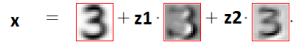
Eigenvalue	2.8162	0.3835
Proportion	0.880	0.120
Cumulative	0.880	1.000

Variable	PC1	PC2
X1	0.523	0.852
X2	0.852	-0.523

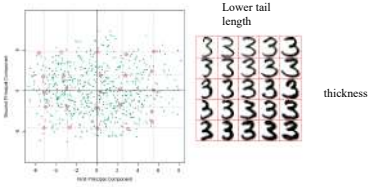
Loadings (U)

Principal Component Analysis

- Digits: two eigenvectors extracted



- Interptretation of eigenvectiors

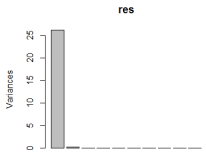


PCA in R

- Prcomp(), biplot(), screeplot()

```
mydata=read.csv2("tecator.csv")
data1=mydata
data1$Fat=c()
res=prcomp(data1)
lambda=res$sdev^2
#eigenvalues
lambda
#proportion of variation
sprintf("%2.3f",lambda/sum(lambda)*100)
screeplot(res)
```

```
> lambda
  [1] 2.612713e+01 2.385369e-01 7.844883e-02 3.018501e-
  [7] 2.052212e-04 1.084213e-04 2.077326e-05 1.150359e-

> sprintf("%2.3f",lambda/sum(lambda)*100)
  [1] "98.679" "0.901"  "0.296"  "0.114"  "0.006
  [9] "0.000"  "0.000"  "0.000"  "0.000"  "0.000
```



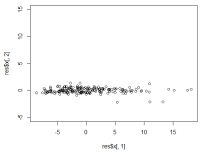
Only 1 component captures the 99% of variation!

PCA in R

- Principal component **loadings (U)**

```
U=res$rotation
head(U)
```

```
> head(U)
              PC1         PC2         PC3
Channel1 0.07938192 0.1156228 0.08073156 -0.0927
Channel2 0.07987445 0.1170972 0.07887873 -0.0981
Channel3 0.08036498 0.1185571 0.07702127 -0.1031
Channel4 0.08085611 0.1200006 0.07515015 -0.1077
Channel5 0.08135022 0.1214075 0.07323819 -0.1119
Channel6 0.08184806 0.1227491 0.07125048 -0.1156
```



- Data in (PC1, PC2) – **scores (Z)**

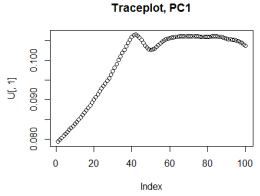
```
plot(res$x[,1], res$x[,2], ylim=c(-5,15))
```

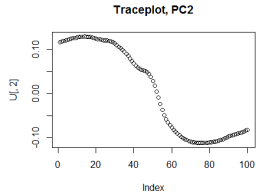
Do we need second dimension?

PCA in R

- Trace plots

```
U= res$rotation
plot(U[,1], main="Traceplot, PC1")
plot(U[,2],main="Traceplot, PC2")
```

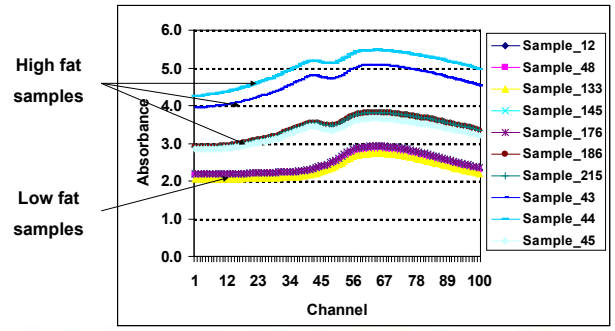




Which components contribute to PC1-2?

Absorbance records for ten samples of chopped meat

PCA2 captures the most of remaining variation



Probabilistic PCA

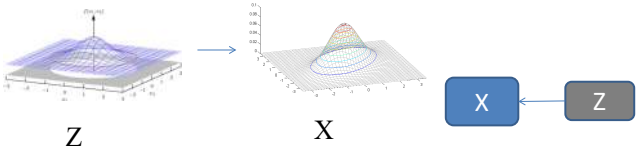
- z_i-latent variables, x_i- observed variables

$$\boldsymbol{z}\sim\boldsymbol{N}(0,\boldsymbol{I})$$
$$\boldsymbol{x}|\boldsymbol{z}\sim\boldsymbol{N}(\boldsymbol{x}|\boldsymbol{Wz}+\boldsymbol{\mu},\sigma^2\boldsymbol{I})$$

- Alternatively

$$\boldsymbol{z}\sim\boldsymbol{N}(0,\boldsymbol{I}),\boldsymbol{x}=\boldsymbol{\mu}+\boldsymbol{Wz}+\boldsymbol{\epsilon},\boldsymbol{\epsilon}\sim\boldsymbol{N}(\boldsymbol{0},\boldsymbol{\sigma^2 I})$$

- **Interpretation**: Observed data (X) is obtained by rotation, scaling and translation of standard normal distribution (Z) and adding some noise.



Probabilistic PCA

- **Aim**: extract Z from X
- Distribution of x:

$$\boldsymbol{x}\sim\boldsymbol{N}(\boldsymbol{\mu},\boldsymbol{C})$$
$$\boldsymbol{C}=\boldsymbol{WW^T}+\boldsymbol{\sigma^2 I}$$

- Rotation invariance
  - Assume that x was generated from $z'=Rz, RR^T=I,\ p(x)$ does not change!

$$\boldsymbol{x}|\boldsymbol{z'}\sim\boldsymbol{N}(\boldsymbol{x}|\boldsymbol{Wz'}+\boldsymbol{\mu},\sigma^2\boldsymbol{I})$$

  - **Model will not be able find latent factors uniquely!** ☹
    - It does not distinguish z from z'

Probabilistic PCA

- Estimation of parameters: ML

Theorem. ML estimates are given by

$$\mu_{ML}=\bar{x}$$
$$W_{ML}=U_M(L_M-\sigma_{ML}^2 I)^{\frac{1}{2}}R$$
$$\sigma_{ML}^2=\frac{1}{p-M}\sum_{i=M+1}^{p}\lambda_i$$

- U_M matrix of M eigenvectors
- L_M diagonal matrix of M eigenvalues
- R any orthogonal matrix

Probabilistic PCA

- Estimation of Z
  - Use mean of posterior

$$\hat{z}=(W_{ML}^T W_{ML}+\sigma_{ML}^2 I)^{-1}W_{ML}^T(x-\mu)$$

- Connection to standard PCA
  - Assume $R=I,\sigma^2=0$→get standard PCA components scaled by inverse root of eigenvalues

$$Z=XUL^{-\frac{1}{2}}$$

Advantages of probabilistic PCA

- More settings to specify→ more flexible
- Can be faster when M<<p
- Missing values can be handled
- M can be derived if a Bayesian version is used
- Probabilistic PCA can be applied to classification problems directly
- Probabilistic PCA can generate new data

Probabilistic PCA in R

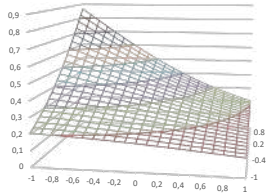
- Use **pcaMethods** from Bioconductor
- Install
  - source("https://bioconductor.org/biocLite.R")
  - biocLite("pcaMethods")

Ppca(data, nPcs,…)

Results: scores, loadings…

Independent component analysis (ICA)

- Probabilistic PCA does not capture latent factors
  - Rotation invariance
- Let's choose distribution which is not rotation invariant→will get unique latent factors
- Choose non-Gaussian $p(z_i)$
- Assuming latent features are **independent**



$$p(z) = \prod_{i=1}^{M} p(z_i) \qquad p(z_i) = \frac{2}{\pi(e^{z_i} + e^{-z_i})}$$

ICA

- Model

$$x = \mu + Wz + \epsilon, \qquad \epsilon \sim \boldsymbol{N}(0, \Sigma)$$

- **Estimation : Maximum likelihood** ($V = W^{-1}$)
  - Assuming noise-free x

$$\max_{V} \sum_{i=1}^{n} \sum_{j=1}^{p} \log\left(p_j(v_j^T x_i)\right)$$

Subject to $||v_i|| = 1$

ICA: estimation algorithm

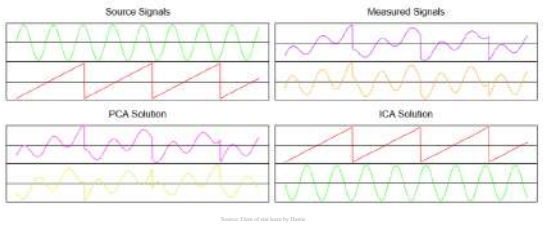
1. Estimate V by maximum likelihood
2. Compute $Z = X'V$

- **With prewhitening**
  1. Convert X into PCA coordinate system (do not remove dimensions): $X' = XU$
  2. Estimate V by maximum likelihood in ICA
  3. Estimate final scores $Z = X'V$

  – Note: full transformation matrix is $U_{ICA} = U \cdot V$

ICA

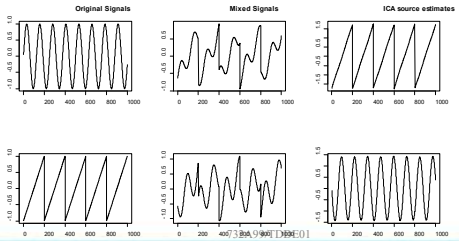
- Example



Independent component analysis: R

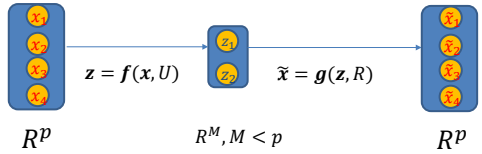
R package: fastICA

```
S <- cbind(sin((1:1000)/20), rep((((1:200)-100)/100), 5))
A <- matrix(c(0.291, 0.6557, -0.5439, 0.5572), 2, 2)
X <- S %*% A #mixing signals
a <- fastICA(X, 2) #now separate them
```



Autoencoders (nonlinear PCA)

- Why linear transformations? Take nonlinear instead!
- $f()$ and $g()$ are typically Neural Networks



$\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{x}, U)$ $\quad \tilde{\boldsymbol{x}} = \boldsymbol{g}(\boldsymbol{z}, R)$

R^p $\quad R^M, M < p$ $\quad R^p$

$$\min_{U,R} \sum_{i=1}^{N} \|x_n - \tilde{x}_n\|^2$$

…or some other loss function

Histogram Classification

- Consider binary classification with input space $\mathbb{R}^D$.
- The best classifier under the 0-1 loss function is $y^*(\boldsymbol{x}) = \arg\max_y p(y|\boldsymbol{x})$.
- Since $\boldsymbol{x}$ may not appear in the finite training set $\{(\boldsymbol{x}_n, t_n)\}$ available, then
  - divide the input space into D-dimensional cubes of side h, and
  - classify according to majority vote in the cube $C(\boldsymbol{x}, h)$ that contains $\boldsymbol{x}$.

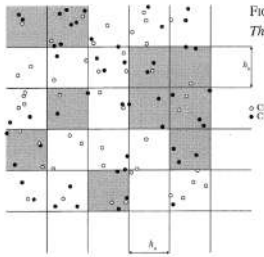


FIGURE 6.1. *A cubic histogram rule: The decision is 1 in the shaded area.*

- In other words,

$$y_C(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \sum_n \mathbf{1}_{\{t_n=1, \boldsymbol{x}_n \in C(\boldsymbol{x},h)\}} \leq \sum_n \mathbf{1}_{\{t_n=0, \boldsymbol{x}_n \in C(\boldsymbol{x},h)\}} \\ 1 & \text{otherwise} \end{cases}$$

Moving Window Classification

- The histogram rule is less accurate at the borders of the cube, because those points are not as well represented by the cube as the ones near the center. Then,
  - consider the points within a certain distance to the point to classify, and
  - classify the point according to majority vote.

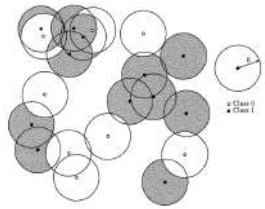


FIGURE 10.1. *The moving window rule in $\mathbb{R}^2$. The decision is 1 in the shaded area.*

- In other words,

$$y_S(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \sum_n \mathbf{1}_{\{t_n=1, \boldsymbol{x}_n \in S(\boldsymbol{x},h)\}} \leq \sum_n \mathbf{1}_{\{t_n=0, \boldsymbol{x}_n \in S(\boldsymbol{x},h)\}} \\ 1 & \text{otherwise} \end{cases}$$

where $S(\boldsymbol{x}, h)$ is a D-dimensional closed ball of radius h centered at $\boldsymbol{x}$.

Kernel Classification

- The moving window rule gives equal weight to all the points in the ball, which may be counterintuitive. Then,

$$y_k(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \sum_n \mathbf{1}_{\{t_n=1\}} k\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right) \leq \sum_n \mathbf{1}_{\{t_n=0\}} k\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right) \\ 1 & \text{otherwise} \end{cases}$$

where $k : \mathbb{R}^D \to \mathbb{R}$ is a kernel function, which is usually non-negative and monotone decreasing along rays starting from the origin. The parameter h is called smoothing factor or width.

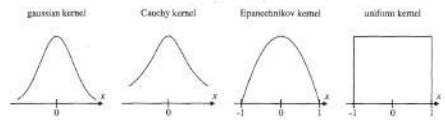


FIGURE 10.3. *Various kernels on $\mathcal{R}$.*

- Gaussian kernel: $k(u) = exp(-\|u\|^2)$ where $\|\cdot\|$ is the Euclidean norm.
- Cauchy kernel: $k(u) = 1/(1 + \|u\|^{D+1})$
- Epanechnikov kernel: $k(u) = (1 - \|u\|^2)\mathbf{1}_{\{\|u\| \leq 1\}}$
- Moving window kernel: $k(u) = \mathbf{1}_{\{u \in S(0,1)\}}$

Kernel Classification

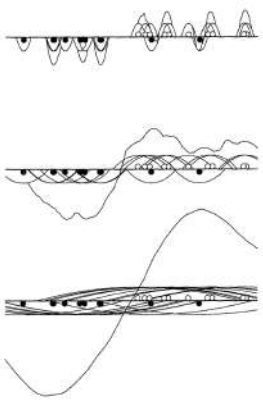


FIGURE 10.2. *Kernel rule on the real line. The figure shows $\sum_{i=1}^n (2Y_i - 1)K((x - X_i)/h)$ for $n = 20$, $K(u) = (1 - u^2)I_{[|u| \leq 1]}$ (the Epanechnikov kernel), and three smoothing factors h. One definitely undersmooths and one oversmooths. We took $p = 1/2$, and the class-conditional densities are $f_0(x) = 2(1 - x)$ and $f_1(x) = 2x$ on $[0, 1]$.*

Histogram, Moving Window, and Kernel Regression

- Consider regressing an unidimensional continuous random variable on a D-dimensional continuous random variable.
- The best regression function under the squared error loss function is $y^*(\boldsymbol{x}) = \mathbb{E}_Y[y|\boldsymbol{x}]$.
- Since $\boldsymbol{x}$ may not appear in the finite training set $\{(\boldsymbol{x}_n, t_n)\}$ available, then we average over the points in $C(\boldsymbol{x}, h)$ or $S(\boldsymbol{x}, h)$, or kernel-weighted average over all the points.
- In other words,

$$y_C(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{x}_n \in C(\boldsymbol{x},h)} t_n}{|\{\boldsymbol{x}_n \in C(\boldsymbol{x}, h)\}|}$$

or

$$y_S(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{x}_n \in S(\boldsymbol{x},h)} t_n}{|\{\boldsymbol{x}_n \in S(\boldsymbol{x}, h)\}|}$$

or

$$y_k(\boldsymbol{x}) = \frac{\sum_n k\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right) t_n}{\sum_n k\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)}$$

Histogram, Moving Window, and Kernel Density Estimation

- Consider density estimation for a D-dimensional continuous random variable.
- Let $R \subseteq \mathbb{R}^D$ and $\boldsymbol{x} \in R$. Then,

$$P = \int_R p(\boldsymbol{x}) d\boldsymbol{x} \simeq p(\boldsymbol{x}) Volume(R)$$

and the number of the N training points $\{\boldsymbol{x}_n\}$ that fall inside R is

$$|\{\boldsymbol{x}_n \in R\}| \simeq P\, N$$

and thus

$$p(\boldsymbol{x}) \simeq \frac{|\{\boldsymbol{x}_n \in R\}|}{N\, Volume(R)}$$

- Then,

$$p_C(\boldsymbol{x}) = \frac{|\{\boldsymbol{x}_n \in C(\boldsymbol{x}, h)\}|}{N\, Volume(C(\boldsymbol{x}, h))}$$

or

$$p_S(\boldsymbol{x}) = \frac{|\{\boldsymbol{x}_n \in S(\boldsymbol{x}, h)\}|}{N\, Volume(S(\boldsymbol{x}, h))}$$

or

$$p_k(\boldsymbol{x}) = \frac{1}{N} \sum_n k\left(\frac{\boldsymbol{x} - \boldsymbol{x}_n}{h}\right)$$

assuming that $k(u) \geq 0$ for all u and $\int k(u) du = 1$.

Histogram, Moving Window, and Kernel Density Estimation

Figure 2.24 An illustration of the histogram approach to density estimation, in which a data set of 50 data points is generated from the distribution shown by the green curve. Histogram density estimates, based on (2.241), with a common bin width Δ are shown for various values of Δ.

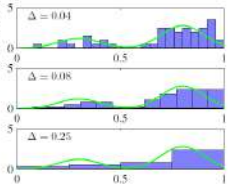
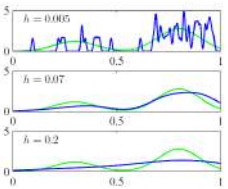


Figure 2.25 Illustration of the kernel density model (2.250) applied to the same data set used to demonstrate the histogram approach in Figure 2.24. We see that h acts as a smoothing parameter and that if it is set too small (top panel), the result is a very noisy density model, whereas if it is set too large (bottom panel), then the bimodal nature of the underlying distribution from which the data is generated (shown by the green curve) is washed out. The best density model is obtained for some intermediate value of h (middle panel).



Histogram, Moving Window, and Kernel Density Estimation

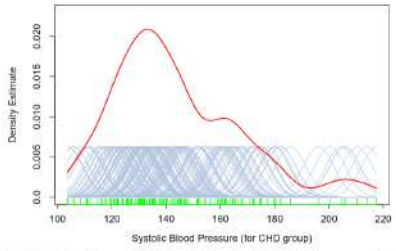


FIGURE 6.13. *A kernel density estimate for systolic blood pressure (for the CHD group). The density estimate at each point is the average contribution from each of the kernels at that point. We have scaled the kernels down by a factor of 10 to make the graph readable.*

Histogram, Moving Window, and Kernel Density Estimation

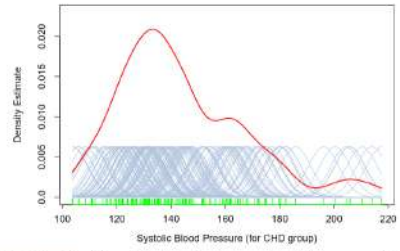


FIGURE 6.13. *A kernel density estimate for systolic blood pressure (for the CHD group). The density estimate at each point is the average contribution from each of the kernels at that point. We have scaled the kernels down by a factor of 10 to make the graph readable.*

- From kernel density estimation to kernel classification:
  1. Estimate $p(\boldsymbol{x}|y = 0)$ and $p(\boldsymbol{x}|y = 1)$ using the methods just seen.
  2. Estimate $p(y)$ as class proportions.
  3. Compute $p(y|\boldsymbol{x}) \propto p(\boldsymbol{x}|y)p(y)$ by Bayes theorem.

Histogram, Moving Window, and Kernel Density Estimation

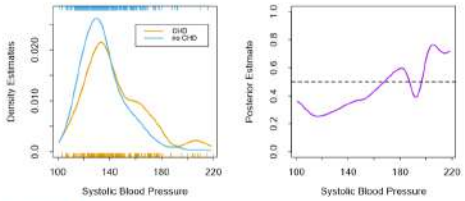
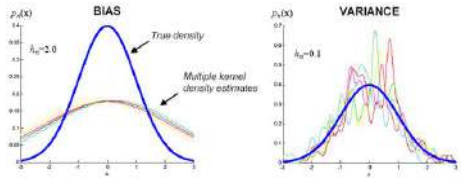


FIGURE 6.14. *The left panel shows the two separate density estimates for systolic blood pressure in the CHD versus no-CHD groups, using a Gaussian kernel density estimate in each. The right panel shows the estimated posterior probabilities for CHD, using (6.25).*

Kernel Selection

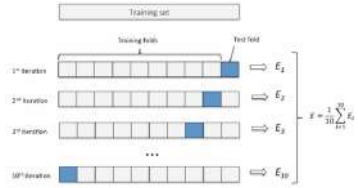
- How to choose the right kernel and width ? E.g., by cross-validation.
- What does "right" mean ? E.g., minimize loss function.
- Note that the width of the kernel corresponds to a bias-variance trade-off.



- Small width implies considering few points. So, the variance will be large (similar to the variance of a single point). The bias will be small since the points considered are close to $\boldsymbol{x}$.
- Large width implies considering many points. So, the variance will be small and the bias will be large.

Kernel Selection

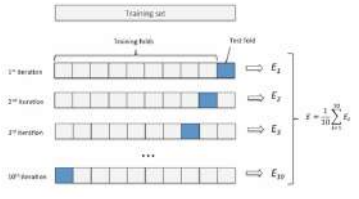
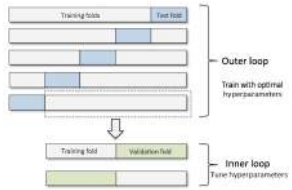
- Recall the following from previous lectures.
- Cross-validation is a technique to estimate the prediction error of a model.



- If the training set contains N points, note that cross-validation estimates the prediction error when the model is trained on $N - N/K$ points.
- Note that the model returned is trained on N points. So, cross-validation overestimates the prediction error of the model returned.
- This seems to suggest that a large K should be preferred. However, this typically implies a large variance of the error estimate, since there are only N/K test points.
- Typically, $K = 5, 10$ works well.

Kernel Selection

- Model: For example, ridge regression with a given value for the penalty factor λ. Only the parameters (weights) need to be determined (closed-form solution).
- Model selection: For example, determine the value for the penalty factor λ. Another example, determine the kernel and width for kernel classification, regression or density estimation. In either case, we do not have a continuous criterion to optimize. Solution: **Nested** cross-validation.

Cross-validation for estimating model prediction error	**Nested** cross-validation for estimating model **selection** prediction error
	

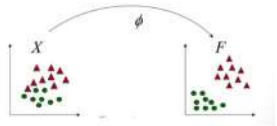
- Error overestimation may not be a concern for model selection. So, $K = 2$ may suffice in the inner loop.
- Which is the fitted model returned by nested cross-validation ?

Kernel Trick

- The kernel function $k\left(\frac{\boldsymbol{x}-\boldsymbol{x}'}{h}\right)$ is invariant to translations, and it can be generalized as $k(\boldsymbol{x}, \boldsymbol{x}')$. For instance,
  - Polynomial kernel: $k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T\boldsymbol{x}' + c)^M$
  - Gaussian kernel: $k(\boldsymbol{x}, \boldsymbol{x}') = exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2\sigma^2)$
- If the matrix

$$\begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_1, \boldsymbol{x}_N) \\ \vdots & \dots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix}$$

is symmetric and positive semi-definite for all choices of $\{\boldsymbol{x}_n\}$, then $k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^T\phi(\boldsymbol{x}')$ where $\phi(\cdot)$ is a mapping from the input space to the feature space.



- The feature space may be non-linear and even infinite dimensional. For instance,

$$\phi(\boldsymbol{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2c}x_1, \sqrt{2c}x_2, c)$$

for the polynomial kernel with $M = D = 2$.

Kernel Trick

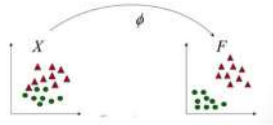
- Consider again moving window classification, regression, and density estimation.
- Note that $\boldsymbol{x}_n \in S(\boldsymbol{x}, h)$ if and only if $\|\boldsymbol{x} - \boldsymbol{x}_n\| \leq h$.
- Note that

$$\|\boldsymbol{x} - \boldsymbol{x}_n\| = \sqrt{(\boldsymbol{x} - \boldsymbol{x}_n)^T(\boldsymbol{x} - \boldsymbol{x}_n)} = \sqrt{\boldsymbol{x}^T\boldsymbol{x} + \boldsymbol{x}_n^T\boldsymbol{x}_n - 2\boldsymbol{x}^T\boldsymbol{x}_n}$$

- Then,

$$\begin{aligned} \|\phi(\boldsymbol{x}) - \phi(\boldsymbol{x}_n)\| &= \sqrt{\phi(\boldsymbol{x}^T)\phi(\boldsymbol{x}) + \phi(\boldsymbol{x}_n^T)\phi(\boldsymbol{x}_n) - 2\phi(\boldsymbol{x}^T)\phi(\boldsymbol{x}_n)} \\ &= \sqrt{k(\boldsymbol{x}, \boldsymbol{x}) + k(\boldsymbol{x}_n, \boldsymbol{x}_n) - 2k(\boldsymbol{x}, \boldsymbol{x}_n)} \end{aligned}$$

- So, the distance is now computed in a (hopefully) more convenient space.



- Note that we do not need to compute $\phi(\boldsymbol{x})$ and $\phi(\boldsymbol{x}_n)$.

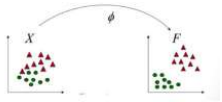
Support Vector Machines for Classification

- Consider binary classification with input space $\mathbb{R}^D$.
- Consider a training set $\{(\boldsymbol{x}_n, t_n)\}$ where $t_n \in \{-1, +1\}$.
- Consider using the linear model

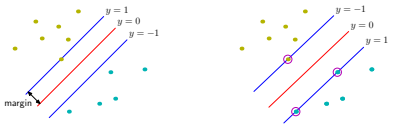
$$y(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x}) + b$$

so that a new point $\boldsymbol{x}$ is classified according to the sign of $y(\boldsymbol{x})$.

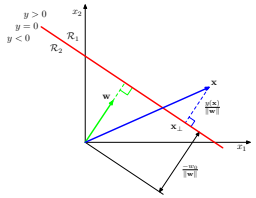
- Assume that the training set is linearly separable in the feature space (but not necessarily in the input space), i.e. $t_n y(\boldsymbol{x}_n) > 0$ for all n.



- Aim for the separating hyperplane that maximizes the margin (i.e. the smallest perpendicular distance from any point to the hyperplane) so as to minimize the generalization error.



Support Vector Machines for Classification



- The perpendicular distance from any point to the hyperplane is given by

$$\frac{t_n y(\boldsymbol{x}_n)}{\|\boldsymbol{w}\|} = \frac{t_n(\boldsymbol{w}^T \phi(\boldsymbol{x}_n) + b)}{\|\boldsymbol{w}\|}$$

- Then, the maximum margin separating hyperplane is given by

$$\arg\max_{\boldsymbol{w},b} \left(\min_n \frac{t_n(\boldsymbol{w}^T \phi(\boldsymbol{x}_n) + b)}{\|\boldsymbol{w}\|} \right)$$

- Multiply $\boldsymbol{w}$ and b by κ so that $t_n(\boldsymbol{w}^T \phi(\boldsymbol{x}_n) + b) = 1$ for the point closest to the hyperplane. Note that $t_n(\boldsymbol{w}^T \phi(\boldsymbol{x}_n) + b)/\|\boldsymbol{w}\|$ does not change.

Support Vector Machines for Classification

- Then, the maximum margin separating hyperplane is given by

$$\arg\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

subject to $t_n(\boldsymbol{w}^T \phi(\boldsymbol{x}_n) + b) \geq 1$ for all n.

- To minimize the previous expression, we minimize

$$\frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_n a_n \big(t_n(\boldsymbol{w}^T \phi(\boldsymbol{x}_n) + b) - 1\big)$$

where $a_n \geq 0$ are called Lagrange multipliers.

- Note that any stationary point of the Lagrangian function is a stationary point of the original function subject to the constraints. Moreover, the Lagrangian function is a quadratic function subject to linear inequality constraints. Then, it is concave, actually concave up because of the +1/2 and, thus, "easy" to minimize.
- Note that we are now minimizing with respect to $\boldsymbol{w}$ and b, and maximizing with respect to a_n.
- Setting its derivatives with respect to $\boldsymbol{w}$ and b to zero gives

$$\boldsymbol{w} = \sum_n a_n t_n \phi(\boldsymbol{x}_n)$$

$$0 = \sum_n a_n t_n$$

Support Vector Machines for Classification

- Replacing the previous expressions in the Lagrangian function gives the dual representation of the problem, in which we maximize

$$\sum_n a_n - \frac{1}{2}\sum_n \sum_m a_n a_m t_n t_m \phi(\boldsymbol{x}_n)^T \phi(\boldsymbol{x}_m) = \sum_n a_n - \frac{1}{2}\sum_n \sum_m a_n a_m t_n t_m k(\boldsymbol{x}_n, \boldsymbol{x}_m)$$

subject to $a_n \geq 0$ for all n, and $\sum_n a_n t_n = 0$.

- Again, this "easy" to maximize.
- Note that the dual representation makes use of the kernel trick, i.e. it allows working in a more convenient feature space without constructing it.

Support Vector Machines for Classification

- When the Lagrangian function is maximized, the Karush-Kuhn-Tucker condition holds for all n:

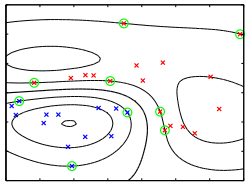
$$a_n(t_n y(\boldsymbol{x}_n) - 1) = 0$$

- Then, $a_n > 0$ if and only if $t_n y(\boldsymbol{x}_n) = 1$. The points with $a_n > 0$ are called support vectors and they lie on the margin boundaries.
- A new point $\boldsymbol{x}$ is classified according to the sign of

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x}) + b = \sum_n a_n t_n \phi(\boldsymbol{x}_n)^T \phi(\boldsymbol{x}) + b = \sum_n a_n t_n k(\boldsymbol{x}, \boldsymbol{x}_n) + b$$

$$= \sum_{m \in \mathcal{S}} a_m t_m k(\boldsymbol{x}, \boldsymbol{x}_m) + b$$

where $\mathcal{S}$ are the indexes of the support vectors. Sparse solution!



Support Vector Machines for Classification

- To find b, consider any support vector $\boldsymbol{x}_n$. Then,

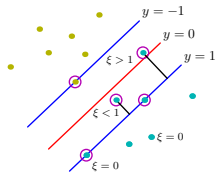
$$1 = t_n y(\boldsymbol{x}_n) = t_n\Big(\sum_{m \in \mathcal{S}} a_m t_m k(\boldsymbol{x}_n, \boldsymbol{x}_m) + b\Big)$$

and multiplying both sides by t_n, we have that

$$b = t_n - \sum_{m \in \mathcal{S}} a_m t_m k(\boldsymbol{x}_n, \boldsymbol{x}_m)$$

- We now drop the assumption of linear separability in the feature space, e.g. to avoid overfitting. We do so by introducing the slack variables $\xi_n \geq 0$ to penalize (almost-)misclassified points as

$$\xi_n = \begin{cases} 0 & \text{if } t_n y(\boldsymbol{x}_n) \geq 1 \\ |t_n - y(\boldsymbol{x}_n)| & \text{otherwise} \end{cases}$$

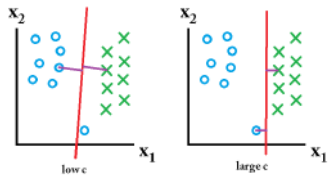


Support Vector Machines for Classification

- The optimal separating hyperplane is given by

$$\arg\min_{\boldsymbol{w},b,\{\xi_n\}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_n \xi_n$$

subject to $t_n y(\boldsymbol{x}_n) \geq 1 - \xi_n$ and $\xi_n \geq 0$ for all n, and where $C > 0$ controls regularization. Its value can be decided by cross-validation. Note that the number of misclassified points is upper bounded by $\sum_n \xi_n$.



- To minimize the previous expression, we minimize

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_n \xi_n - \sum_n a_n\big(t_n(\boldsymbol{w}^T \phi(\boldsymbol{x}_n) + b) - 1 + \xi_n\big) - \sum_n \mu_n \xi_n$$

where $a_n \geq 0$ and $\mu_n \geq 0$ are Lagrange multipliers.

Support Vector Machines for Classification

- Setting its derivatives with respect to $\boldsymbol{w}$, b and ξ_n to zero gives

$$\boldsymbol{w} = \sum_n a_n t_n \phi(\boldsymbol{x}_n)$$

$$0 = \sum_n a_n t_n$$

$$a_n = C - \mu_n$$

- Replacing these in the Lagrangian function gives the dual representation of the problem, in which we maximize

$$\sum_n a_n - \frac{1}{2}\sum_n \sum_m a_n a_m t_n t_m k(\boldsymbol{x}_n, \boldsymbol{x}_m)$$

subject to $a_n \geq 0$ and $a_n \leq C$ for all n, because $\mu_n \geq 0$.

- When the Lagrangian function is maximized, the Karush-Kuhn-Tucker conditions hold for all n:

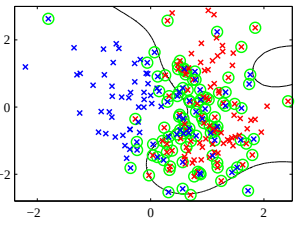
$$a_n(t_n y(\boldsymbol{x}_n) - 1 + \xi_n) = 0$$

$$\mu_n \xi_n = 0$$

- Then, $a_n > 0$ if and only if $t_n y(\boldsymbol{x}_n) = 1 - \xi_n$ for all n. The points with $a_n > 0$ are called support vectors and they lie
  - on the margin if $a_n < C$, because then $\mu_n > 0$ and thus $\xi_n = 0$, or
  - inside the margin (even on the wrong side of the decision boundary) if $a_n = C$, because then $\mu_n = 0$ and thus ξ_n is unconstrained.

Support Vector Machines for Classification

- Since the optimal $\boldsymbol{w}$ takes the same form as in the linearly separable case, classifying a new point is done the same as before. Finding b is done the same as before by considering any support vector $\boldsymbol{x}_n$ with $0 < a_n < C$.



- Not covered topics:
  - Classifying into more than two classes.
  - Returning class posterior probabilities.

Support Vector Machines for Regression

- Consider regressing an unidimensional continuous random variable on a D-dimensional continuous random variable.
- Consider a training set $\{(\boldsymbol{x}_n, t_n)\}$. Consider using the linear model

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x}) + b$$

- To get a sparse solution, instead of minimizing the classical regularized error function

$$\frac{1}{2}\sum_n (y(\boldsymbol{x}_n) - t_n)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

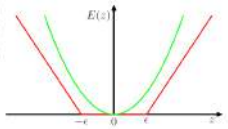
consider minimizing the ϵ-insensitive regularized error function

$$C\sum_n E_\epsilon(y(\boldsymbol{x}_n) - t_n) + \frac{1}{2}\|\boldsymbol{w}\|^2$$

where $C > 0$ controls regularization and

$$E_\epsilon(y(\boldsymbol{x}) - t) = \begin{cases} 0 & \text{if } |y(\boldsymbol{x}) - t| < \epsilon \\ |y(\boldsymbol{x}) - t| - \epsilon & \text{otherwise} \end{cases}$$

Figure 7.6 Plot of an ϵ-insensitive error function (in red) in which the error increases linearly with distance beyond the insensitive region. Also shown for comparison is the quadratic error function (in green).



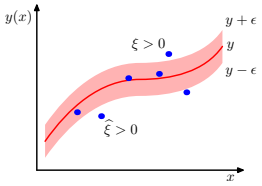
Support Vector Machines for Regression

- The values of C and ϵ can be decided by cross-validation.
- Consider the slack variables $\xi_n \geq 0$ and $\widehat{\xi}_n \geq 0$ such that

$$\xi_n = \begin{cases} t_n - y(\boldsymbol{x}_n) - \epsilon & \text{if } t_n > y(\boldsymbol{x}_n) + \epsilon \\ 0 & \text{otherwise} \end{cases}$$

and

$$\widehat{\xi}_n = \begin{cases} y(\boldsymbol{x}_n) - \epsilon - t_n & \text{if } t_n < y(\boldsymbol{x}_n) - \epsilon \\ 0 & \text{otherwise} \end{cases}$$



Support Vector Machines for Regression

- The optimal regression curve is given by

$$\arg\min_{\boldsymbol{w}, b, \{\xi_n\}, \{\widehat{\xi}_n\}} C\sum_n (\xi_n + \widehat{\xi}_n) + \frac{1}{2}\|\boldsymbol{w}\|^2$$

subject to $\xi \geq 0$, $\widehat{\xi}_n \geq 0$, $t_n \leq y(\boldsymbol{x}_n) + \epsilon + \xi_n$ and $t_n \geq y(\boldsymbol{x}_n) - \epsilon - \widehat{\xi}_n$.

- To minimize the previous expression, we minimize

$$C\sum_n (\xi_n + \widehat{\xi}_n) + \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_n (\mu_n \xi_n + \widehat{\mu}_n \widehat{\xi}_n)$$

$$- \sum_n a_n (y(\boldsymbol{x}_n) + \epsilon + \xi_n - t_n) - \sum_n \widehat{a}_n (t_n - y(\boldsymbol{x}_n) + \epsilon + \widehat{\xi}_n)$$

where $\mu_n \geq 0$, $\widehat{\mu}_n \geq 0$, $a_n \geq 0$ and $\widehat{a}_n \geq 0$ are Lagrange multipliers.

- Setting its derivatives with respect to $\boldsymbol{w}$, b, ξ_n and $\widehat{\xi}_n$ to zero gives

$$\boldsymbol{w} = \sum_n (a_n - \widehat{a}_n)\phi(\boldsymbol{x}_n)$$
$$0 = \sum_n (a_n - \widehat{a}_n)$$
$$C = a_n + \mu_n$$
$$C = \widehat{a}_n + \widehat{\mu}_n$$

Support Vector Machines for Regression

- Replacing these in the Lagrangian function gives the dual representation of the problem, in which we maximize

$$\frac{1}{2}\sum_n \sum_m (a_n - \widehat{a}_n)(a_m - \widehat{a}_m) k(\boldsymbol{x}_n, \boldsymbol{x}_m) - \epsilon \sum_n (a_n + \widehat{a}_n) + \sum_n (a_n - \widehat{a}_n) t_n$$

subject to $a_n \geq 0$ and $a_n \leq C$ for all n, because $\mu_n \geq 0$. Similarly for $\widehat{a}_n$.

- When the Lagrangian function is maximized, the Karush-Kuhn-Tucker conditions hold for all n:

$$a_n(y(\boldsymbol{x}_n) + \epsilon + \xi_n - t_n) = 0$$
$$\widehat{a}_n(t_n - y(\boldsymbol{x}_n) + \epsilon + \widehat{\xi}_n) = 0$$
$$\mu_n \xi_n = 0$$
$$\widehat{\mu}_n \widehat{\xi}_n = 0$$

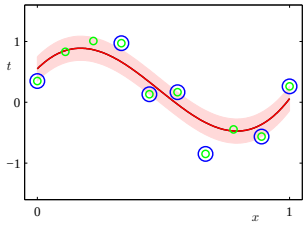
- Then, $a_n > 0$ if and only if $y(\boldsymbol{x}_n) + \epsilon + \xi_n - t_n = 0$, which implies that $\boldsymbol{x}_n$ lies on or above the upper margin of the ϵ-tube. Similarly for $\widehat{a}_n > 0$.

Support Vector Machines for Regression

- The prediction for a new point $\boldsymbol{x}$ is made according to

$$y(\boldsymbol{x}) = \sum_{m \in \mathcal{S}} (a_m - \widehat{a}_m) k(\boldsymbol{x}, \boldsymbol{x}_m) + b$$

where $\mathcal{S}$ are the indexes of the support vectors. Sparse solution!



- To find b, consider any support vector $\boldsymbol{x}_n$ with $0 < a_n < C$. Then, $\mu_n > 0$ and thus $\xi_n = 0$ and thus $0 = t_n - \epsilon - y(\boldsymbol{x}_n)$. Then,

$$b = t_n - \epsilon - \sum_{m \in \mathcal{S}} (a_m - \widehat{a}_m) k(\boldsymbol{x}_n, \boldsymbol{x}_m)$$

Neural Networks

- Consider binary classification with input space $\mathbb{R}^D$. Consider a training set $\{(\boldsymbol{x}_n, t_n)\}$ where $t_n \in \{-1, +1\}$.
- SVMs classify a new point $\boldsymbol{x}$ according to

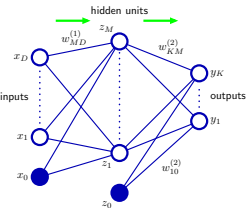
$$y(\boldsymbol{x}) = \text{sgn}\Big(\sum_{m\in\mathcal{S}} a_m t_m k(\boldsymbol{x}, \boldsymbol{x}_m) + b\Big)$$

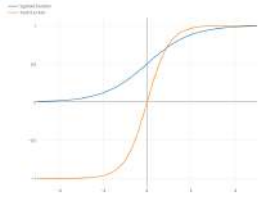
- Consider regressing an unidimensional continuous random variable on a D-dimensional continuous random variable. Consider a training set $\{(\boldsymbol{x}_n, t_n)\}$
- For a new point $\boldsymbol{x}$, SVMs predict

$$y(\boldsymbol{x}) = \sum_{m\in\mathcal{S}} (a_n - \widehat{a}_n) k(\boldsymbol{x}, \boldsymbol{x}_m) + b$$

- SVMs imply data-selected user-defined basis functions.
- NNs imply a user-defined number of data-selected basis functions.

Neural Networks





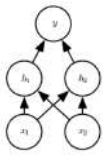
- Activations: $a_j = \sum_i w_{ji}^{(1)} x_i + w_{j0}^{(1)}$
- Hidden units and activation function: $z_j = h(a_j)$
- Output activations: $a_k = \sum_j w_{kj}^{(2)} z_j + w_{k0}^{(2)}$
- Output activation function for regression: $y_k(\boldsymbol{x}) = a_k$
- Output activation function for classification: $y_k(\boldsymbol{x}) = \sigma(a_k)$
- Sigmoid function: $\sigma(a) = \frac{1}{1+\exp(-a)}$
- Two-layer NN:

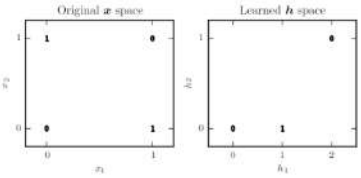
$$y_k(\boldsymbol{x}) = \sigma\Big(\sum_j w_{kj}^{(2)} h\Big(\sum_i w_{ji}^{(1)} x_i + w_{j0}^{(1)}\Big) + w_{k0}^{(2)}\Big)$$

- Evaluating the previous expression is known as forward propagation. The NN is said to have a feed-forward architecture.
- All the previous is, of course, generalizable to more layers.

Neural Networks

- Solving the XOR problem with NNs.
- No line shatters the points in the original space.
- The NN represents a mapping of the input space to an alternative space where a line can shatter the points. Note that the points (0,1) and (1,0) are mapped both to the point (1,0).
- **It resembles SVMs**.





$w_{11}^{(1)} = w_{12}^{(1)} = w_{21}^{(1)} = w_{22}^{(1)} = 1$
$w_{10}^{(1)} = 0,\ w_{20}^{(1)} = -1$
$h_j = z_j = h(a_j) = \max\{0, a_j\}$
$w_{11}^{(2)} = 1,\ w_{12}^{(2)} = -2$
$w_{10}^{(2)} = 0$
$y = y_k = a_k$

Backpropagation Algorithm

- Consider regressing an K-dimensional continuous random variable on a D-dimensional continuous random variable.
- Consider a training set $\{(\boldsymbol{x}_n, \boldsymbol{t}_n)\}$. Consider minimizing the sum-of-squares error function

$$E(\boldsymbol{w}) = \sum_n E_n(\boldsymbol{w}) = \sum_n \frac{1}{2}\|\boldsymbol{y}(\boldsymbol{x}_n) - \boldsymbol{t}_n\|^2 = \sum_n \sum_k \frac{1}{2}(y_k(\boldsymbol{x}_n) - t_{nk})^2$$

- This error function can be justified from a maximum likelihood approach to learning $\boldsymbol{w}$. To see it, **assume** that

$$p(t_k|\boldsymbol{x}, \boldsymbol{w}, \sigma) = \mathcal{N}(t_k|y_k(\boldsymbol{x}), \sigma)$$

- Then, the likelihood function is

$$p(\{\boldsymbol{t}_n\}|\{\boldsymbol{x}_n\}, \boldsymbol{w}, \sigma) = \prod_n \prod_k \mathcal{N}(t_{nk}|y_k(\boldsymbol{x}_n), \sigma) = \prod_n \prod_k \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(t_{nk} - y_k(\boldsymbol{x}_n))^2}$$

and thus

$$-\ln p(\{\boldsymbol{t}_n\}|\{\boldsymbol{x}_n\}, \boldsymbol{w}, \sigma) = \sum_n \sum_k \frac{1}{2\sigma^2}(t_{nk} - y_k(\boldsymbol{x}_n))^2 + \frac{N}{2}\ln\sigma^2 + \frac{N}{2}\ln 2\pi$$

which is equivalent to the sum-of-squares error function for a given σ.

- If σ is not given, then we can find the ML estimates of $\boldsymbol{w}$, plug them into the log likelihood function, and maximize it with respect to σ.

Backpropagation Algorithm

- The weight space is highly multimodal and, thus, we have to resort to approximate iterative methods to minimize the previous expression.
- Batch gradient descent

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta_t \nabla E(\boldsymbol{w}^t)$$

where $\eta_t > 0$ is the learning rate ($\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$ to ensure convergence, e.g. $\eta_t = 1/t$).

- Sequential, stochastic or online gradient descent

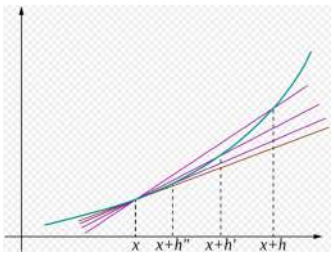
$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta_t \nabla E_n(\boldsymbol{w}^t)$$

where n is chosen randomly or sequentially.

- Sequential gradient descent is less affected by the multimodality problem, as a local minimum of the whole data will not be generally a local minimum of each individual point.

Backpropagation Algorithm

- Recall that $f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}$

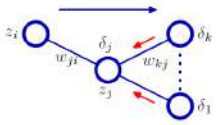


- Recall that $\nabla E_n(\boldsymbol{w}^t)$ is a vector whose components are the partial derivatives of $E_n(\boldsymbol{w}^t)$.

Backpropagation Algorithm

- Backpropagation algorithm:
  1. Forward propagate to compute activations, and hidden and output units.
  2. Compute δ_k for the output units.
  3. Backpropagate the δ's, i.e. evaluate δ_j for the hidden units recursively.
  4. Compute the required derivatives.

Figure 5.7 Illustration of the calculation of δ_j for hidden unit j by backpropagation of the δ's from those units k to which unit j sends connections. The blue arrow denotes the direction of information flow during forward propagation, and the red arrows indicate the backward propagation of error information.



- For classification, we minimize the negative log likelihood function, a.k.a. cross-entropy error function:

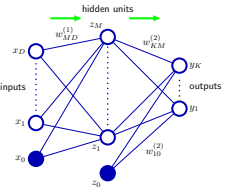
$$E_n(\boldsymbol{w}) = -\sum_k [t_{nk} \ln y_k(\boldsymbol{x}_n) + (1 - t_{nk}) \ln(1 - y_k(\boldsymbol{x}_n))]$$

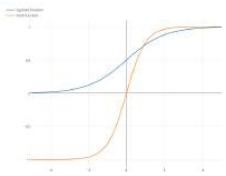
with $t_{nk} \in \{0, 1\}$ and $y_k(\boldsymbol{x}_n) = \sigma(a_k)$. Then, again

$$\frac{\partial E_n}{\partial w_{kj}} = \delta_k z_j \text{ and } \delta_k = \frac{\partial E_n}{\partial a_k} = y_k - t_k$$

- This is an example of embarrassingly parallel algorithm.

Backpropagation Algorithm





- Example: $y_k = a_k$, and $z_j = h(a_j) = \tanh(a_j)$ where $\tanh(a) = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}$.
- Note that $h'(a) = 1 - h(a)^2$.
- Backpropagation:
  1. Forward propagation, i.e. compute
  $$a_j = \sum_i w_{ji} x_i \text{ and } z_j = h(a_j) \text{ and } y_k = \sum_j w_{kj} z_j$$
  2. Compute
  $$\delta_k = y_k - t_k$$
  3. Backpropagate, i.e. compute
  $$\delta_j = (1 - z_j^2) \sum_k w_{kj} \delta_k$$
  4. Compute
  $$\frac{\partial E_n}{\partial w_{kj}} = \delta_k z_j \text{ and } \frac{\partial E_n}{\partial w_{ji}} = \delta_j x_i$$

Backpropagation Algorithm

- The weight space is non-convex and has many symmetries, plateaus and local minima. So, the initialization of the weights in the backpropagation algorithm is crucial.
- Hints based on experimental rather than theoretical analysis:
  - Initialize the weights to different values, otherwise they would be updated in the same way because the algorithm is deterministic, and so creating redundant hidden units.
  - Initialize the weights at random, but
    - too small magnitude values may cause losing signal in the forward or backward passes, and
    - too big magnitude values may cause the activation function to saturate and lose gradient.
  - Initialize the weights according to prior knowledge: Almost-zero for hidden units that are unlikely to interact, and bigger magnitude values for the rest.
  - Initialize the weights to almost-zero values so that the initial model is almost-linear, i.e. the sigmoid function is almost-linear around the zero. Let the algorithm to introduce non-linearities where needed.
    - Note however that this initialization makes the sigmoid function take a value around half its saturation level. That is why the hyperbolic tangent function is sometimes preferred in practice.

Regularization

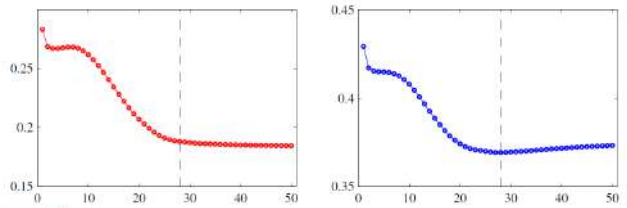


Figure 5.12 An illustration of the behaviour of training set error (left) and validation set error (right) during a typical training session, as a function of the iteration step, for the sinusoidal data set. The goal of achieving the best generalization performance suggests that training should be stopped at the point shown by the vertical dashed lines, corresponding to the minimum of the validation set error.

- Regularization when learning the parameters: Early stopping the backpropagation algorithm according to the error on some validation data.
- Regularization when learning the structure:
  - Cross-validation.
  - Penalizing complexity according to

$$E(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \text{ or } E(\boldsymbol{w}) + \frac{\lambda_1}{2}\|\boldsymbol{w}^{(1)}\|^2 + \frac{\lambda_2}{2}\|\boldsymbol{w}^{(2)}\|^2$$

and choose λ, or λ_1 and λ_2 by cross-validation. Note that the effect of the penalty is simply to add λw_{ji} and λw_{kj}, or $\lambda_1 w_{ji}$ and $\lambda_2 w_{kj}$ to the appropriate derivatives.

Limitations of Neural Networks

Theorem (Universal approximation theorem)

For every continuous function $f : [a, b]^D \to \mathbb{R}$ and for every $\epsilon > 0$, there exists a NN with one hidden layer such that

$$\sup_{\mathbf{x} \in [a,b]^D} |f(\mathbf{x}) - y(\mathbf{x})| < \epsilon$$

Theorem (Universal classification theorem)

Let $C^{(k)}$ contain all classifiers defined by NNs of one hidden layer with k hidden units and the sigmoid activation function. Then, for any distribution $p(\mathbf{x}, t)$,

$$\lim_{k \to \infty} \inf_{y \in C^{(k)}} L(y(\mathbf{x})) - L(p(t|\mathbf{x})) = 0$$

where $L()$ is the 0/1 loss function.

- How many hidden units has such a NN ?
- How much data do we need to learn such a NN (and avoid overfitting) via the backpropagation algorithm ?
- How fast does the backpropagation algorithm converge to such a NN ? Assuming that it does not get trapped in a local minimum...
- The answer to the last two questions depends on the first: More hidden units implies more training time and higher generalization error.

Limitations of Neural Networks

- How many hidden units does the NN need ?
- Any Boolean function can be written in disjunctive normal form (OR of ANDs) or conjunctive normal form (AND of ORs). This is a depth-two logical circuit.
- For most Boolean functions, the size of the circuit is exponential in the size of the input.
- However, there are Boolean functions that have a polynomial-size circuit of depth k and an exponential-size circuit of depth $k - 1$.
- Then, there is no universally right depth. Ideally, we should let the data determine the right depth.

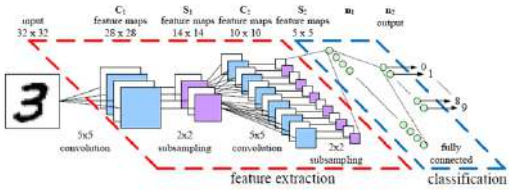
Theorem (No free lunch theorem)

For any algorithm, good performance on some problems comes at the expense of bad performance on some others.

Deep Neural Networks

- Training DNNs is difficult:
  - Typically, poorer generalization than (shallow) NNs.
  - The gradient may vanish/explode as we move away from the output layer, due to multiplying small/big quantities. E.g. the gradient of σ and tanh is in $[0, 1)$. So, they may only suffer the gradient vanishing problem. Other activations functions may suffer the gradient exploding problem.
  - There may be larger plateaus and many more local minima than with NNs.
- Training DNNs is doable:
  - Convolutional networks, particularly suitable for image processing.
  - Rectifier activation function, a new activation function.
  - Layer-wise pre-training, to find a good starting point for training.
- In addition to performance, the computational demands of the training must be considered, e.g. CPU, GPU, memory, parallelism, etc.
  - The authors state that GoogLeNet was trained "using modest amount of model and data-parallelism. Although we used a CPU based implementation only, a rough estimate suggests that the GoogLeNet network could be trained to convergence using few high-end GPUs within a week, the main limitation being the memory usage".

Convolutional Networks



- DNNs suitable for image recognition, since they exhibit invariance to translation, scaling, rotations, and warping.
- Convolution: Detection of local features, e.g. a_j is computed from a 5x5 pixel patch of the image.
- To achieve invariance, the units in the convolution layer share the same activation function and weights.
- Subsampling: Combination of local features into higher-order features, e.g. a_k is compute from a 2x2 pixel patch of the convoluted image.
- There are several feature maps in each layer, to compensate the reduction in resolution by increasing in the number of features being detected.
- The final layer is a regular NN for classification.

Convolutional Networks

- DNNs allow increased depth because
  - they are sparse, which allows the gradient to propagate further, and
  - they have relatively few weights to fit due to feature locality and weight sharing.
- The backpropagation algorithm needs to be adapted, by modifying the derivatives with respect to the weights in each convolution layer m.
- Since E_n depends on $w_i^{(m)}$ only via $a_j^{(m)}$, and $a_j^{(m)} = \sum_{i \in L_j^{(m)}} w_i^{(m)} z_i^{(m-1)}$ where $L_j^{(m)}$ is the set of indexes of the input units, then

$$\frac{\partial E_n}{\partial w_i^{(m)}} = \sum_j \frac{\partial E_n}{\partial a_j^{(m)}} \frac{\partial a_j^{(m)}}{\partial w_i^{(m)}} = \sum_j \delta_j^{(m)} z_i^{(m-1)}$$

- Note that $w_i^{(m)}$ does not depend on j by weight sharing, whereas $i \in L_j^{(m)}$ by feature locality.

Rectifier Activation Function

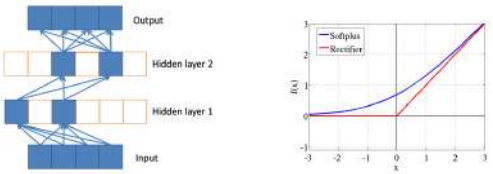


Figure 2: *Left:* **Sparse propagation of activations and gradients in a network of rectifier units.** The input selects a subset of active neurons and computation is linear in this subset. *Right:* **Rectifier and softplus activation functions.** The second one is a smooth version of the first.

- It simplifies the backpropagation algorithm as $h'(a_j) = 1$ for the selected units. So, there is no gradient vanishing on the paths of selected units. Compare with the sigmoid or hyperbolic tangent, for which
  - the gradient is smaller than one, or
  - even zero due to saturation.
- Note that $h'(0)$ does not exist since $h'_+(0) \neq h'_-(0)$. We can get around this problem by simply returning one of two one-sided derivatives. Or using a generalization of the rectifier function.
- Regularization is typically added to prevent numerical problems due to the activation being unbounded, e.g. when forward propagating.

Rectifier Activation Function

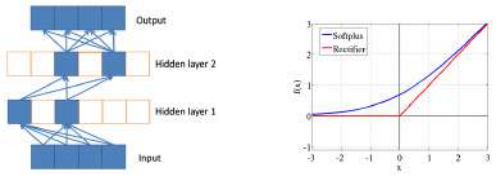


Figure 2: *Left:* **Sparse propagation of activations and gradients in a network of rectifier units.** The input selects a subset of active neurons and computation is linear in this subset. *Right:* **Rectifier and softplus activation functions.** The second one is a smooth version of the first.

- *rectifier*$(x) = \max\{0, x\}$, i.e. hidden units are off or operating in a linear regime.
- The most popular choice nowadays.
- Sparsity promoting: Uniform initialization of the weights implies that around 50 % of the hidden units are off.
- Piece-wise linear mapping: The input selects which hidden units are active, and the output is a liner function of the input in the selected hidden units.

Innehåll

Regression

Linear regression

Estimation: maximum likelihood is equal to min square: $RSS(w) = \sum_{i=1}^{n} (Y_i - \boldsymbol{w}^T \boldsymbol{X}_i)^2$. Optimal condition: $X^T (y - Xw) = 0$

X has 1s as first col: cbind(1,X), y is one col. Gives estimation as: w_hat = $(X^T X)^{-1} * X^T y$.

R: w_hat/beta = solve(t(X)%*%X, t(X)%*%Y)

Package: lm().

Things

Residuals, difference between obs value vs model prediction. Use residuals() in R. Error.

R commands

Pic rows/col

X[1,6] X[1:9,] X[-(1:9),] Entire row 4: X[4,] Pic row, col 5>: X[X[,5]>20,]

Transpose: T(X)

Inverse: solve(X) $d=X^{-1}$ b = solve(X,b)

Packages used in lab

- readr
- glmnet
- kknn
- MASS
- tree or (rpart) I choose tree.
- e1071 - Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier
- fastICA
- geosphere
- neuralnet

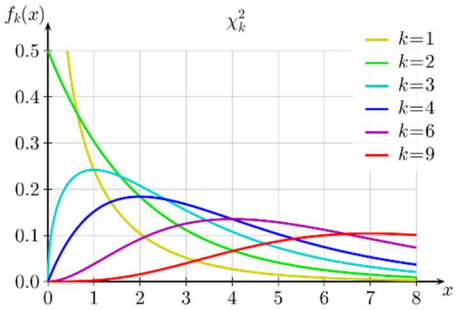
extra assign:

Distributions

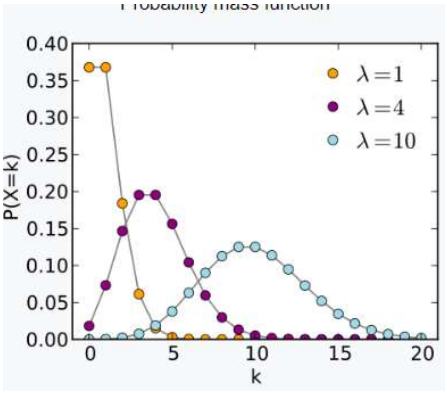
Chi-square (två) distribution

Chitvåfördelning alternativt Chikvadratfördelning, χ^2-fördelning, är inom matematisk statistik en kontinuerlig sannolikhetsfördelning med täthetsfunktionen:

$$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

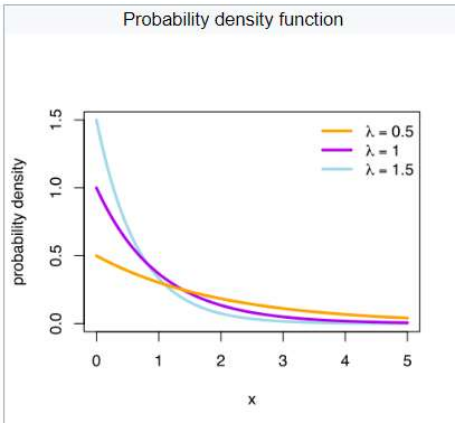


Poisson distribution



$$\frac{\lambda^k e^{-\lambda}}{k!}$$

Exponential distribution



$$\lambda e^{-\lambda x}$$

Labs

Lab1

Assignment 1 - Spam classification with nearest neighbors

```
library(readr)
library(glmnet)
data <- read_csv2("C:/Users/oskar/OneDrive/Universitet/Linköping Universitet/År4/Machine learning/Lab 1/spambase.csv")

n=dim(data)[1]
set.seed(12345)
id=sample(1:n, floor(n*0.5))
train=data[id,]
test=data[-id,]

predModel = glm(Spam~., data=train, family = binomial)

missclass=function(X, Xfit){
  n=length(X)
  print(table(X, Xfit))
  return (1-sum(diag(table(X, Xfit)))/n)
}

#Pred 1
prediction_test = predict(predModel, test, type="response")
prediction2_test = prediction_test
prediction_test[prediction_test<=0.5]=0
prediction_test[prediction_test>0.5]=1
plot(prediction_test)
missClassResult_test = missclass(test[[49]], prediction_test)
print(missClassResult_test)

prediction_train = predict(predModel, train, type="response")
prediction2_train = prediction_train
prediction_train[prediction_train<=0.5]=0
prediction_train[prediction_train>0.5]=1
plot(prediction_train)
missClassResult_train = missclass(train[[49]], prediction_train)
print(missClassResult_train)

#Pred 2
prediction2_train[prediction2_train<=0.8]=0
prediction2_train[prediction2_train>0.8]=1

prediction2_test[prediction2_test<=0.8]=0
prediction2_test[prediction2_test>0.8]=1
#plot(prediction2)

missClassResult_8_train = missclass(train[[49]], prediction2_train)
missClassResult_8_test = missclass(test[[49]], prediction2_test)
print(missClassResult_8_train)
print(missClassResult_8_test) #New rule makes more missclassifications.
```

```
#kknn
library(kknn)

knnn= kknn(as.factor(Spam)~., train, test, k=30)
missClassResultk30 = missclass(test[[49]], knnn$fitted.values)
knnn= kknn(as.factor(Spam)~., train, train, k=30)
missClassResultk30_train = missclass(train[[49]], knnn$fitted.values)

knnn1= kknn(as.factor(Spam)~., train, test, k=1)
missClassResultk1 = missclass(test[[49]], knnn1$fitted.values)
knnn1= kknn(as.factor(Spam)~., train, train, k=1)
missClassResultk1_train = missclass(train[[49]], knnn1$fitted.values)

print(missClassResultk30)
print(missClassResultk30_train)

print(missClassResultk1) #not as good
print(missClassResultk1_train)
```

Assignment 2 - Inference about lifetime of machines

```
machines<- read_csv2("C:/Users/oskar/OneDrive/Universitet/Linköping Universitet/År4/Machine learning/Lab 1/machines.csv")

hist(machines$Length)#exponential

#log-likelihood
log_likelihood = function(theta, data){
  p=0
  data = data$Length
  for(i in data){
    p = p + log( theta*exp(-theta*i))
  }
  return (p)
}

#find my Theta
findTheta = function(data){
  p = 1:100
  i=1
  for(theta in seq(from=0, to=10, by=0.1)){
    p[i]= log_likelihood(theta, data)
    i = i+1
  }
  theta = seq(from=0, to=10, by=0.1)
  plot(theta, p)
  return(p)
}

logPTotal = findTheta(machines)
```

```
logPSix = findTheta(machines[(1:6), ])

thetaTotal = seq(from=0, to=10, by=0.1)[which.max(logPTotal)]
thetaSixFirst = seq(from=0, to=10, by=0.1)[which.max(logPSix)]

print(thetaTotal)
print(thetaSixFirst)
theta = seq(from=0, to=10, by=0.1)
plot(theta, logPTotal, col="blue", ylim=c(-100,0))
par(new=TRUE)
plot(theta, logPSix, col="red", ylim=c(-100,0))

#whti is this?
#curve(dim(machines)[1]*log(x)-x*sum(machines), from=min(machines), ylim=c(-80,0), col="blue",  to=4,
ylab="log(p(x|??))", sub="Red: 6 obs | Blue: All obs", xlab="??", add=FALSE)

#part4
#log-likelihood Bayesian
#log_likelihood_Bayesian = function(theta,lambda, data){
#  p=0
#  data = data$Length
#  for(i in data){
#    p = p +  #log( theta*exp(-theta*i)*lambda*exp(-lambda*theta))
#  }
#  return (p)
#}

findThetaBayesian = function(lambda, data){
  p = 1:100
  i=1
  for(theta in seq(from=0, to=10, by=0.1)){
    log_like_lambda = log(lambda)+(-theta*lambda)
    p[i]= log_like_lambda + log_likelihood(theta, data)
    i = i+1
  }
  theta = seq(from=0, to=10, by=0.1)
  plot(theta, p)
  return(p)
}

theta = seq(from=0, to=10, by=0.1)
LogsThetaBay = findThetaBayesian(10, machines)
thetaMaxBay = seq(from=0, to=10, by=0.1)[which.max(LogsThetaBay)]

plot(theta, logPTotal, col="blue", ylim=c(-100,-30))
par(new=TRUE)
plot(theta, LogsThetaBay, ylim=c(-100,-30))
print(thetaMaxBay)



#5
set.seed(12345)
#thetaTotal = seq(from=0, to=10, by=0.1)[which.max(logPTotal)]
```

```
new_Data = rexp(50, rate=thetaTotal)
old_Data = machines$Length

#plot hist in one diagram
p1 <- hist(old_Data)
p2 <- hist(new_Data)
plot( p1, col=rgb(0,0,1,1/4), xlim=c(0,7), ylim=c(0,35))  # first histogram
plot( p2, col=rgb(1,0,0,1/4), xlim=c(0,7),ylim=c(0,35), add=T) #second histogram

# Both behave the same way. Are distributed alike, both following the exponential distributon.
# new data followingtheta=1.1. which are generated from old_data
```

Assignment 3 - Feature selection by cross-validation in a linear model - Extra

```
#linear regression and returns predicted Y
mylin=function(X,Y, Xpred){
  Xpred1=cbind(1,Xpred)
 #MISSING: check formulas for linear regression and compute beta
  #minimizing least sqeare givew following formula: w_hat = (Xt*X)^-1 * Xt * y
  X = cbind(1, X)
  beta = solve(t(X)%*%X, t(X)%*%Y)
  Res=Xpred1%*%beta
  return(Res)
}


myCV=function(X,Y,Nfolds){
  n=length(Y)
  p=ncol(X)
  set.seed(12345)
  ind=sample(n,n)
  X1=X[ind,]
  Y1=Y[ind]
  sF=floor(n/Nfolds)
  MSE=numeric(2^p-1)
  Nfeat=numeric(2^p-1)
  Features=list()
  curr=0

  #we assume 5 features.

  for (f1 in 0:1)
    for (f2 in 0:1)
      for(f3 in 0:1)
        for(f4 in 0:1)
          for(f5 in 0:1){
            model= c(f1,f2,f3,f4,f5)
            if (sum(model)==0) next()
            SSE=0

            for (k in 1:Nfolds){
              #MISSING: compute which indices should belong to current fold
```

```
        if(k!=Nfolds){
          indices = ((k-1)*sF):(k*sF)
        }else{
          indices = ((k-1)*sF):n
        }

        X_train = X1[-indices, which(model == 1)]
        Y_train = Y1[-indices]
        X_validate = X1[indices, which(model == 1)]
        Yp = Y1[indices]

                                                    #MISSING: implement cross-validation for model with features in
"model" and iteration i.
                                                    #MISSING: Get the predicted values for fold 'k', Ypred, and the
original values for folf 'k', Yp.
        Ypred = mylin(X_train, Y_train, X_validate)

         SSE=SSE+sum((Ypred-Yp)^2)
       }
       curr=curr+1
       MSE[curr]=SSE/n
       Nfeat[curr]=sum(model)
       Features[[curr]]=model

     }

 #MISSING: plot MSE against number of features
 plot(Nfeat, MSE, main = "MSE", xlab = "Number of features")

 i=which.min(MSE)
 return(list(CV=MSE[i], Features=Features[[i]]))
}

myCV(as.matrix(swiss[,2:6]), swiss[[1]], 5)
```

Assignment 4 - Linear regression and regularization

```
tecator <- read_csv2("C:/Users/oskar/OneDrive/Universitet/Linköping Universitet/År4/Machine learning/Lab
1/tecator.csv")

#-------1---------
plot(tecator$Moisture, tecator$Protein)
# Looks lika a line, makes me think that a linear model would be good.

#-------2---------
#Consider model ???????????????? in which Moisture is normally distributed, and the expected
#Moisture is a polynomial function of Protein including the polynomial terms up to power
#???????? (i.e M1 is a linear model, M2 is a quadratic model and so on). Report a probabilistic
#model that describes ????????????????. Why is it appropriate to use MSE criterion when fitting
#this model to a training data?
```

```
#-------3---------

n=dim(tecator)[1]
set.seed(12345)
id=sample(1:n, floor(n*0.5))
t_train=tecator[id,]
t_test=tecator[-id,]

mse = function(x, x_pred){
  squared_error = (x-x_pred)^2
  mse = sum(squared_error)/length(x)
  return(mse)
}

#sum (i=0-n) bi*x^i

model_i_1 = glm(Moisture~Protein, data=t_train)
model_i_2 = glm(Moisture~Protein + I(Protein^2), data=t_train)
model_i_3 = glm(Moisture~Protein + I(Protein^2) + I(Protein^3), data=t_train)
model_i_4 = glm(Moisture~Protein + I(Protein^2) + I(Protein^3) + I(Protein^4), data=t_train)
model_i_5 = glm(Moisture~Protein + I(Protein^2) + I(Protein^3) + I(Protein^4) + I(Protein^5), data=t_train)
model_i_6 = glm(Moisture~Protein + I(Protein^2) + I(Protein^3) + I(Protein^4) + I(Protein^5) + I(Protein^6),
data=t_train)

#predictions for test data
pred_i_1 = predict(model_i_1, newdata = t_test)
pred_i_2 = predict(model_i_2, newdata = t_test)
pred_i_3 = predict(model_i_3, newdata = t_test)
pred_i_4 = predict(model_i_4, newdata = t_test)
pred_i_5 = predict(model_i_5, newdata = t_test)
pred_i_6 = predict(model_i_6, newdata = t_test)

#predictions for train data
pred_i_1_train = predict(model_i_1, newdata = t_train)
pred_i_2_train = predict(model_i_2, newdata = t_train)
pred_i_3_train = predict(model_i_3, newdata = t_train)
pred_i_4_train = predict(model_i_4, newdata = t_train)
pred_i_5_train = predict(model_i_5, newdata = t_train)
pred_i_6_train = predict(model_i_6, newdata = t_train)

#mse for test data
mse_test_1 = mse(t_test$Moisture, pred_i_1)
mse_test_2 = mse(t_test$Moisture, pred_i_2)
mse_test_3 = mse(t_test$Moisture, pred_i_3)
mse_test_4 = mse(t_test$Moisture, pred_i_4)
mse_test_5 = mse(t_test$Moisture, pred_i_5)
mse_test_6 = mse(t_test$Moisture, pred_i_6)

#mse for train data
mse_test_1_train = mse(t_train$Moisture, pred_i_1_train)
mse_test_2_train = mse(t_train$Moisture, pred_i_2_train)
mse_test_3_train = mse(t_train$Moisture, pred_i_3_train)
mse_test_4_train = mse(t_train$Moisture, pred_i_4_train)
```

```
mse_test_5_train = mse(t_train$Moisture, pred_i_5_train)
mse_test_6_train = mse(t_train$Moisture, pred_i_6_train)

print(mse_test_1)
print(mse_test_2)
print(mse_test_3)
print(mse_test_4)
print(mse_test_5)
print(mse_test_6)

print(mse_test_1_train)
print(mse_test_2_train)
print(mse_test_3_train)
print(mse_test_4_train)
print(mse_test_5_train)
print(mse_test_6_train)

mses_test = c(mse_test_1, mse_test_2, mse_test_3, mse_test_4, mse_test_5, mse_test_6)
mses_train = c(mse_test_1_train, mse_test_2_train, mse_test_3_train, mse_test_4_train, mse_test_5_train,
mse_test_6_train)
which.min(mses_test)
which.min(mses_train)

plot((1:6), mses_test, col="red")

plot((1:6), mses_train, col="red")

# Both
plot((1:6), mses_test, col="blue", ylim=c(23,45) )
par(new=TRUE)
plot((1:6), mses_train, col="red", ylim=c(23,45)  )



#4
sub_of_tecator = subset(tecator, select = - c(Protein, Moisture, Sample))

n=dim(sub_of_tecator)[1]
set.seed(12345)
id=sample(1:n, floor(n*0.5))
t_train_sub=sub_of_tecator[id,]
t_test_sub=sub_of_tecator[-id,]

library(MASS)
linear_model_fat = glm(Fat~., data=sub_of_tecator)
step = stepAIC(linear_model_fat, direction = "both")

step$anova
summary(step)

covariates = scale(subset(sub_of_tecator, select = -Fat))
response = scale(subset(sub_of_tecator, select = Fat))

#-------5---------ridge
```

```
ridge = glmnet(as.matrix(covariates), response, alpha=0, family= "gaussian" )
plot(ridge, xvar="lambda", label=TRUE)

#-------6---------
lasso = glmnet(as.matrix(covariates), response, alpha=1, family= "gaussian")
plot(lasso, xvar = "lambda", label=TRUE) #higher lambda, lower vairance, higher variance.

#-------7---------
#gamma = 1, gives me lasso(0 is ridge), lambda = is which lambda to include.
lasso_mse = cv.glmnet(covariates, response, gamma=1, lambda = seq(from = 0, to=1, by=0.01))
plot(lasso_mse)


#library #i=3 best.(readr)
#spambase <- read_delim("C:/Users/oskar/OneDrive/Universitet/Linköping Universitet/År4/Machine learning/Lab
1/spambase.csv",
#                       ";", escape_double = FALSE, trim_ws = TRUE)
```

Lab 2

Assignment 1 - LDA and logistic regression

```
RNGversion('3.5.1')
library(readr)
set.seed(12345)

#Assignment1
australian_crabs = read.csv("C:/Users/oskar/OneDrive/Universitet/Linköping Universitet/År4/Machine learning/Lab
2/australian-crabs.csv")

#--------step1----------
australian_crabs_males = subset(australian_crabs, sex=="Male")
australian_crabs_females = subset(australian_crabs, sex=="Female")

plot(australian_crabs_males[['CL']], australian_crabs_males[['RW']], ylim=c(6,20), xlim=c(15,45), col="red",
ylab="RW", xlab="CL")
par(new=TRUE)
plot(australian_crabs_females[['CL']], australian_crabs_females[['RW']], ylim=c(6,20), xlim=c(15,45), col="blue",
ylab="RW", xlab="CL")


#---------Step2----------
library(MASS)

lda_pred = lda(sex~CL + RW, data=australian_crabs)

print(lda_pred)
pred = predict(lda_pred, australian_crabs)
table(australian_crabs[['sex']], pred$class)

predicted_dataset = data.frame(pred$class, australian_crabs[['CL']], australian_crabs[['RW']])
names(predicted_dataset) = c('sex', 'CL', 'RW')
```

```
plot(subset(predicted_dataset, sex=="Male")[['CL']], subset(predicted_dataset, sex=="Male")[['RW']], ylim=c(6,20),
xlim=c(15,45), col="red", ylab="RW", xlab="CL")
par(new=TRUE)
plot(subset(predicted_dataset, sex=="Female")[['CL']], subset(predicted_dataset, sex=="Female")[['RW']],
ylim=c(6,20), xlim=c(15,45), col="blue", ylab="RW", xlab="CL")

# Misclassification function
misclass=function(X, Xfit){
  n=length(X)
  return (1-sum(diag(table(X, Xfit)))/n)
}
lda_pred_misclassification = misclass(australian_crabs[['sex']], pred$class)
print(lda_pred_misclassification)

#-------step3-------
lda_pred_wprior = lda(sex~CL + RW, data=australian_crabs, prior = c(0.1, 0.9))
print(lda_pred_wprior)

pred_wprior = predict(lda_pred_wprior, australian_crabs)
table(australian_crabs[['sex']], pred_wprior$class)
predicted_dataset_wprior = data.frame(pred_wprior$class, australian_crabs[['CL']], australian_crabs[['RW']])
names(predicted_dataset_wprior) = c('sex', 'CL', 'RW')

plot(subset(predicted_dataset_wprior, sex=="Male")[['CL']], subset(predicted_dataset_wprior, sex=="Male")[['RW']],
ylim=c(6,20), xlim=c(15,45), col="red", ylab="RW", xlab="CL")
par(new=TRUE)
plot(subset(predicted_dataset_wprior, sex=="Female")[['CL']], subset(predicted_dataset_wprior,
sex=="Female")[['RW']], ylim=c(6,20), xlim=c(15,45), col="blue", ylab="RW", xlab="CL")

lda_pred_misclassification_wprior = misclass(australian_crabs[['sex']], pred_wprior$class)
print(lda_pred_misclassification_wprior)

#--------step 4----------
#check if sex is as factor
str(australian_crabs)

logistic_regression = glm(as.factor(sex) ~ CL + RW, data=australian_crabs, family = binomial)
print(logistic_regression)

#dessission boundary
intercept = coef(logistic_regression)[1]/(-coef(logistic_regression)[3])
slope = coef(logistic_regression)[2]/(-coef(logistic_regression)[3])
x = seq(15,45, by=1)
y = slope*x + intercept


prediction_LR = predict(logistic_regression, australian_crabs, type="response")
predicted_sex = prediction_LR
predicted_sex[prediction_LR<0.5] = 'Female' #prediction_LR will return all rows that are under threashold!
predicted_sex[prediction_LR>=0.5] = 'Male'

predicted_sex
australian_crabs[["sex"]]
```

```
misclass(australian_crabs[["sex"]], predicted_sex)

table(australian_crabs[["sex"]], predicted_sex)

predicted_dataset_LR = data.frame(predicted_sex, australian_crabs[['CL']], australian_crabs[['RW']])
names(predicted_dataset_LR) = c('sex', 'CL', 'RW')


plot(x, y, type="l", ylim=c(6,20),  xlim=c(15,45), ylab="RW", xlab="CL")
par(new=TRUE)
plot(subset(predicted_dataset_LR, sex=="Male")[['CL']], subset(predicted_dataset_LR, sex=="Male")[['RW']],
ylim=c(6,20), xlim=c(15,45), col="red", ylab="RW", xlab="CL")
par(new=TRUE)
plot(subset(predicted_dataset_LR, sex=="Female")[['CL']], subset(predicted_dataset_LR, sex=="Female")[['RW']],
ylim=c(6,20), xlim=c(15,45), col="blue", ylab="RW", xlab="CL")
```

Assignment 2 - Analysis of credit scoring

```
#----------Step 1----------
creditscoring = read.csv2("C:/Users/oskar/OneDrive/Universitet/Linköping Universitet/År4/Machine learning/Lab
2/creditscoring.csv")

RNGversion('3.5.1')

n=dim(creditscoring)[1]
set.seed(12345)
id=sample(1:n, floor(n*0.5))
train=creditscoring[id,]

id1=setdiff(1:n, id)
set.seed(12345)
id2=sample(id1, floor(n*0.25))
valid=creditscoring[id2,]

id3=setdiff(id1,id2)
test=creditscoring[id3,]

library(tree)
#or:
library(rpart)

#Training data to fit model
fit_deviance = tree(good_bad~. , split = "deviance", data = train)
fit_gini = tree(good_bad~. , split = "gini", data = train)

summary(fit_deviance)
summary(fit_gini)

#Predict using test data.
predict_deviance = predict(fit_deviance, newdata = test, type = "class")

#table(test[["good_bad"]], predict_deviance)
```

```
misclass_deviance = misclass(test[["good_bad"]], predict_deviance)
print(misclass_deviance)

predict_gini = predict(fit_gini, newdata = test, type = "class")

#table(test[["good_bad"]], predict_gini)
misclass_gini = misclass(test[["good_bad"]], predict_gini)
print(misclass_gini)

#----------Step 3----------
#Deviance is chosen due to lower misclassification rate for test data.
summary(fit_deviance)

train_score = rep(0,15)
test_score = rep(0,15)

for(i in 2:15) {
  pruned_tree = prune.tree(fit_deviance, best = i)
  pred = predict(pruned_tree, newdata=valid, type="tree")

  train_score[i] = deviance(pruned_tree)
  test_score[i] = deviance(pred)
}

plot(2:15, train_score[2:15], type="b", col="red", ylim=c(200,550), ylab="Deviance", xlab="No. of leaves")
points(2:15, test_score[2:15], type="b", col="blue")

test_score[1] = 5000
which.min(test_score)

## Min when best=4
test_score[4]
pruned_tree = prune.tree(fit_deviance, best = 4)
summary(pruned_tree)
plot(pruned_tree)
text(pruned_tree, pretty = 0)

#Misclass for test
prediction_test = predict(pruned_tree, newdata = test, type = "class")
table(test[["good_bad"]], prediction_test)
misclass(test[["good_bad"]], prediction_test)

#----------Step 4 ----------
library(MASS)
library(e1071)

fit_naive_bayes =naiveBayes(good_bad~., data=train)
summary(fit_naive_bayes)
#train data
predict_naive_bayes_train = predict(fit_naive_bayes, newdata = train)
table(train[["good_bad"]], predict_naive_bayes_train)
misclass(train[["good_bad"]], predict_naive_bayes_train)
#test data
```

```
predict_naive_bayes_test = predict(fit_naive_bayes, newdata = test)
table(test[["good_bad"]], predict_naive_bayes_test)
misclass(test[["good_bad"]], predict_naive_bayes_test)
# remember: 1-(sum(diag(table))/sum(table))

#----------Step 5-----------
# TPR = true positive rate(y-axis)
# FPR = false positive reate(x-axis)
predict_naive_bayes_test = predict(fit_naive_bayes, newdata = test, type= "raw")
predict_naive_bayes_test

pi = seq(from = 0.05, to = 0.95, by = 0.05 )
n = length(pi)

#Naive Bayes
TPR = rep(0,n)
FPR = rep(0,n)
for( i in 1:n){
  predict = predict_naive_bayes_test[,2]
  predict = ifelse(predict>pi[i], "good", "bad")
  table = table(test[["good_bad"]], predict)
  print(table)
  TPR[i] = (table[2, 2])/sum(table[2, ])
  FPR[i] = (table[1, 2])/sum(table[1, ])
}

# tree ROC
#str(test)
prediction_test = predict(pruned_tree, newdata = test, type = "vector")

n = length(pi)
TPR_tree = rep(0,n)
FPR_tree = rep(0,n)
for( i in 1:n){
  pred = as.vector(prediction_test[,2])
  pred = ifelse(pred>pi[i], "good", "bad")
  if ( sum(pred=="bad")==0) {
    FPR_tree[i] = 1
    TPR_tree[i] = 1
  } else if ( sum(pred=="good")==0) {
    TPR_tree[i] = 0
    FPR_tree[i] = 0
  } else {
    table = table(test[["good_bad"]], pred)
    print(table)
    TPR_tree[i] = (table[2, 2])/sum(table[2, ])
    FPR_tree[i] = (table[1, 2])/sum(table[1, ])
  }
}

plot(FPR_tree, TPR_tree, xlim = (0:1), ylim= (0:1), type="b", col="red", xlab="FPR", ylab="TPR", main="ROC")
par(new=TRUE)
plot(FPR, TPR, xlim = (0:1), ylim= (0:1), type="b", col="blue", xlab="FPR", ylab="TPR")
```

```
#-------Step 6-------------
fit_naive_bayes =naiveBayes(good_bad~., data=train)
summary(fit_naive_bayes)
#train data
naive_bayes_train = predict(fit_naive_bayes, newdata = train, type="raw")
predict_train = ifelse(naive_bayes_train[,2]/naive_bayes_train[,1]>10, "good", "bad")

table(train[["good_bad"]], predict_train)
misclass(train[["good_bad"]], predict_train)

#test data
naive_bayes_test = predict(fit_naive_bayes, newdata = test, type="raw")
predict_test = ifelse(naive_bayes_test[,2]/naive_bayes_test[,1]>10, "good", "bad")
misclass(test[["good_bad"]], predict_test)
table = table(test[["good_bad"]], predict_test)
print(table)
```

Assignment 3 - Uncertainty estimation - Extra

```
RNGversion('3.5.1')
library(readr)

state = read.csv2("C:/Users/oskar/OneDrive/Universitet/Linköping Universitet/År4/Machine learning/Lab
2/state.csv")
state = data.frame(state)
#---Step 1---
state_ordered = state[order(state$MET, decreasing = FALSE),]
plot(state_ordered$MET, state_ordered$EX, main ="MET vs EX", xlab = "MET", ylab = "EX")
# qudratic or polinomial?

#---Step 2---
library(tree)
tree_model = tree(EX ~ MET, state_ordered, control = tree.control(nobs = nrow(state_ordered), minsize = 8))
plot(tree_model)
text(tree_model, pretty = 0)
cv_tree = cv.tree(tree_model)
best_size = cv_tree$size[which.min(cv_tree$dev)]

#following plot showes that best = 4 is the best tree.
plot(x=cv_tree$size, y=cv_tree$dev, type = "b", col= "blue")

pruned_tree = prune.tree(tree_model, best = best_size)
pred = predict(pruned_tree, newdata = state_ordered)

hist(residuals(pruned_tree), breaks = 20)

plot(x=state_ordered$MET, y=state_ordered$EX)
points(x=state_ordered$MET , y=pred, col="red", type = "l")
```

```
#---step 3---
set.seed(12345)
B=1000
sample_size = nrow(state_ordered)

boot_predictions = matrix(nrow=sample_size, ncol=B)
for(i in 1:B){
  samples = sample(seq(1,nrow(state_ordered),1), size = sample_size, replace=TRUE)
  data = state_ordered[samples, ]

  tree_model = tree(EX ~ MET, data = data, control = tree.control(nobs= nrow(state_ordered), minsize = 8))
  pruned_tree = prune.tree(tree_model, best=best_size)
  boot_predictions[, i] = predict(pruned_tree, newdata = state_ordered)
}
#calcultate std deviations and confidence bands using quantiles function
conf_band = apply(boot_predictions, 1, function(x){
  band = quantile(x, probs = c(0.025, 0.975))
  return(band)
})

#Plot conf. intervalls:
plot(x=state_ordered$MET, y=state_ordered$EX, col = "red", xlab = "MET", ylab = "EX", main = "95% conf. bands")
points(x=state_ordered$MET , y=pred, col="red", type = "l")
lines(x=state_ordered$MET, y=conf_band[1, ], col = "blue")
lines(x=state_ordered$MET, y=conf_band[2, ], col = "blue")


#use bootstrap ----------TEST
library(boot)
#function to generate datapoints for non-parametic bootstrap
f=function(data, ind){
  data1=data[ind,]# extract bootstrap sample
  tree_model=tree(EX ~ MET, data1, control = tree.control(nobs = nrow(data1), minsize = 8)) #fit linear model
  pruned_tree = prune.tree(tree_model, best = best_size)

  #predict values for all Area values from the original data
  priceP=predict(pruned_tree, newdata=state_ordered)
  return(priceP)
}
res=boot(state_ordered, f, R=1000) #make bootstrap
res
boot_data = data.frame(res$t)
conf_band = apply(boot_data, 2, function(col){
  band = quantile(col, probs = c(0.025, 0.975))
  return(band)})
conf_band

plot(x=state_ordered$MET, y=state_ordered$EX, col = "red", xlab = "MET", ylab = "EX", main = "95% conf. bands")
points(x=state_ordered$MET , y=pred, col="red", type = "l")
lines(x=state_ordered$MET, y=conf_band[1, ], col = "blue")
lines(x=state_ordered$MET, y=conf_band[2, ], col = "blue")

# end ----------TEST
```

```
#-----Step 4-------
#parametric bootstrap for confidence intervalls:
#values for distr_gen. sigmna = std_dev
tree_model = tree(EX ~ MET, state_ordered, control = tree.control(nobs = nrow(state_ordered), minsize = 8))
pruned_tree = prune.tree(tree_model, best = best_size)
pred = predict(pruned_tree, newdata = state_ordered)
residuals = state_ordered$EX - pred
std_dev = sd(residuals)

distr_gen = function(data, tree_model){
  pred = predict(tree_model, newdata = data)
  res = rnorm(nrow(data), mean = pred, sd= std_dev )
  data$EX = res
  return(data)
}

tree_model = pruned_tree

stat = function(data){
 tree_model = tree(EX ~ MET, data=data,
             control = tree.control(nobs = nrow(data),
                          minsize = 8))
 pruned_tree = prune.tree(tree_model, best = best_size)
 pred = predict(pruned_tree, newdata = state_ordered)
 return(pred)
}

res_para = boot(state_ordered, statistic = stat ,
   mle = tree_model,
   R = 1000,
   sim = "parametric",
   ran.gen = distr_gen )

boot_data = data.frame(res_para$t)
conf_band = apply(boot_data, 2, function(col){
 band = quantile(col, probs = c(0.025, 0.975))
 return(band)})
conf_band

plot(x=state_ordered$MET, y=state_ordered$EX, col = "red", xlab = "MET", ylab = "EX", main = "95% conf. bands")
points(x=state_ordered$MET , y=pred, col="red", type = "l")
lines(x=state_ordered$MET, y=conf_band[1, ], col = "blue")
lines(x=state_ordered$MET, y=conf_band[2, ], col = "blue")

test = envelope(res_para)
lines(x=state_ordered$MET, y=test$point[1,], col = "green")
lines(x=state_ordered$MET, y=test$point[2,], col = "green")

#prediction bands
statistic = function(data){
 tree_model = tree(EX ~ MET, data=data,
             control = tree.control(nobs = nrow(data),
```

```
                                minsize = 8))
  pruned_tree = prune.tree(tree_model, best = best_size)
  pred = predict(pruned_tree, newdata = state_ordered)
  res = rnorm(nrow(data), mean = pred, sd= std_dev )
  return(res)
}
res_pred_band = boot(state_ordered, statistic = statistic ,
            mle = tree_model,
            R = 1000,
            sim = "parametric",
            ran.gen = distr_gen )
pred_band = envelope(res_pred_band)
lines(x=state_ordered$MET, y=pred_band$point[1,])
lines(x=state_ordered$MET, y=pred_band$point[2,])

#step 5
hist(residuals(pruned_tree),
    breaks = 20, main ="Histogram of the residuals",
    xlab = "Residual")
#chi-square model would be preffered.
```

Assignment 4 - Principal components

```
NIR_spectra = read.csv2("C:/Users/oskar/OneDrive/Universitet/Linköping Universitet/År4/Machine learning/Lab
2/NIRSpectra.csv")

#--------Step 1------------
data1 = NIR_spectra
data1$Viscosity = c()
res = prcomp(data1)

#squaring sdev to get values that are (proportional to) eigenvalues
lambda = res$sdev^2
X = res$x

#hom much variance is explained in each component
sprintf("%2.3f",lambda/sum(lambda)*100)

#histogram of explained variance
screeplot(res)

# extract 2 components to get 99 explenation of total variance. PC1, PC2.
plot(res$x[,1], res$x[,2], xlab ="PC1", ylab="PC2")

#----------Step 2---------------
plot(res$rotation[,1], main="Traceplot of PC1")
plot(res$rotation[,2], main="Traceplot of PC2")

#----------Step 3-------------
library(fastICA)
set.seed(12345)
ica = fastICA(data1, 2)
```

```
W_fnutt = ica$K %*% ica$W
plot(W_fnutt[,1], main="Traceplot of W'1")
plot(W_fnutt[,2], main="Traceplot of W'2")

#Plot of scores for the two latent features
plot(ica$S, main="ICA Score", xlab="Latent Feature 1", ylab="Latent Feature 2")

#TESTing
plot(ica$X, main = "Pre-processed data")
plot(ica$X %*% ica$K, main = "PCA components")
plot(ica$S, main = "ICA components")
plot(ica$K)
```

Lab 3

Assignment 1 - KERNEL METHODS

```
RNGversion('3.5.1')
library(readr)
library(geosphere)

#---Assignment 1 ----
stations = read.csv("C:/Users/oskar/OneDrive/Universitet/Linköping Universitet/År4/Machine learning/Lab 3/stations.csv")
temps = read.csv("C:/Users/oskar/OneDrive/Universitet/Linköping Universitet/År4/Machine learning/Lab 3/temps50k.csv")

set.seed(1234567890)
st <- merge(stations,temps,by="station_number")


h_distance <- 80000 # These three values are up to the students
h_date <- 10
h_time <- 4
a <- 58.4274 # The point to predict (up to the students)
b <- 14.826
date <- "2013-11-04" # The date to predict (up to the students)
times <- c("04:00:00", "06:00:00", "08:00:00", "10:00:00", "12:00:00", "14:00:00", "16:00:00", "18:00:00", "20:00:00", "22:00:00", "24:00:00")
temp <- vector(length=length(times))
temp_mult <- vector(length=length(times))

# Students' code here
#test h values
max(distHaversine(data.frame(st$latitude, st$longitude), c(a,b)))
dist = seq(0,200000, 1)
y = exp(-(dist/h_distance)^2)
plot(y, type="l", main = "Distance kernel")
#satisfied with h_distance = 80000 beacause then we stop to care if distance>than200km
dat = seq(0,30, 1)
y = exp(-(dat/h_date)^2)
plot(y, type="l", main = "Date kernel")
#satisfied with h_date = 10 because then we stop to care if the date is older than 25 days.
tim = seq(0, 24, 1)
```

```
y = exp(-(tim/h_time)^2)
plot(y, type="l", main = "Time kernel")
#satisfied with h_time = 4 because then only times within 7 hours is used for estimate.

# remove all posterior dates
str(st)
st$date = as.Date(st$date, format = "%Y-%m-%d")
#remove earlier times than 04:00:00
filtered_data = st[st$date <= date, ]
filtered_data = filtered_data[!(filtered_data$date==date && substr(filtered_data$time, 1, 2)<substr(times[1], 1, 2))]

#Gussian kernel is used: k(u) = exp(-||u||^2),
#||.|| is the Euclidean norm.
euclidean = function(X){
  return (sqrt(sum(X^2)))
}

#gussian kernel
kernel_dist = function(X, X_n, h){
  distance = distHaversine(X, X_n)
  u = (distance)/h  # calculate u = X-Xn/h
  return(exp(-euclidean(u))) #calculate k
}

kernel_date = function(X, X_n, h){
  distance = as.numeric((X - X_n))%%365.25
  if (distance > 365/2) {
    distance = 365 - distance
  }
  u = distance / h
  return(exp(-euclidean(as.numeric(u))))
}

kernel_time = function(X, X_n, h){
  distance = as.numeric(X) - as.numeric(X_n)
  if (distance > 12){
    distance = 24-distance
  }
  u = distance/h
  return(exp(-euclidean(u)))
}



#calculate y with dist
#filtered_data = filtered_data[order(filtered_data$latitude, filtered_data$longitude),]

#kernel_weight = 0
n = nrow(filtered_data)
k=vector("numeric", length = n)
k_mult = vector("numeric", length = n)
k_loop=vector("numeric", length = n)
k_mult_loop = vector("numeric", length = n)
```

```
for(i in 1:n){
  k_dist = kernel_dist(c(filtered_data$longitude[i], filtered_data$latitude[i]), c(a,b), h_distance )
  k_date = kernel_date(filtered_data$date[i], as.Date(date), h_date)

  k[i]=k_date + k_dist
  k_mult[i] = k_date * k_dist
}
for(j in 1:(length(times))){
  for (i in 1:n) {
    k_time = kernel_time(substr(filtered_data$time[i], 1, 2), substr(times[j], 1, 2), h_time)
    k_loop[i] = k[i] + k_time
    k_mult_loop[i] = k_mult[i] * k_time

    temp[j] = temp[j] + k_loop[i]*filtered_data$air_temperature[i]
    temp_mult[j] = temp_mult[j] + k_mult_loop[i] * filtered_data$air_temperature[i]
  }
  temp[j] = temp[j] / sum(k_loop)
  temp_mult[j] = temp_mult[j] /sum(k_mult_loop)
}

#test_temp = test_temp/kernel_weight #kernel weighted temp.
#test_temp
plot(temp, xaxt = "n", type = "b", main = "Kernel Addition")
axis(1, at=1:length(df.times), labels=df.times)

plot(temp_mult, xaxt = "n", type = "b", main ="Kernel multiplication")
axis(1, at=1:length(df.times), labels=df.times)
```

Assignment 2 - SUPPORT VECTOR MACHINES - Extra

```
##Use the function ksvm from the R package kernlab to learn a SVM for classifying the spam dataset that is included
#with the package. Consider the radial basis function kernel (also known as Gaussian) with a width of 0.05. For the
#parameter C, consider values 0.5, 1 and 5. This implies that you have to consider three models.
# Perform model selection, i.e. select the most promising of the three models (use any method of your choice except
#cross-validation or nested-cross-validation)
# Estimate the generalization error of the SVM selected above (use any method of your choice except cross-
validation
#or nested cross validation)
# Produce the SVM that will be returned to the user, i.e. show the code
# What is the purpose of the parameter C?

library(kernlab)
set.seed(1234567890)
data(spam)

#Create function for misclassification rate
missclass=function(conf_matrix, fit_matrix){
  n=length(fit_matrix[,1])
  return(1-sum(diag(conf_matrix))/n)
}


index=sample(1:4601)
```

```
train=spam[index[1:2500],]
valid=spam[index[2501:3501],]
test=spam[index[3502:4601],]

svmmodel1=ksvm(type~., data=train, kernel="rbfdot", kpar=list(sigma=0.05), C=0.5)
pred1=predict(svmmodel1, newdata=valid)
confusion1=table(valid$type, pred1)
misclass1=missclass(confusion1, valid)
print(confusion1)
print(misclass1)

svmmodel2=ksvm(type~., data=train, kernel="rbfdot", kpar=list(sigma=0.05), C=1)
pred2=predict(svmmodel2, newdata=valid)
confusion2=table(valid$type, pred2)
misclass2=missclass(confusion2, valid)
print(confusion2)
print(misclass2)

svmmodel3=ksvm(type~., data=train, kernel="rbfdot", kpar=list(sigma=0.05), C=5)
pred2=predict(svmmodel3, newdata=valid)
confusion3=table(valid$type, pred2)
misclass3=missclass(confusion3, valid)
print(confusion3)
print(misclass3)

##Conclusion: The model with the C value of 1 is the best since it has the lowest misclassification rate. However,
##since the application is classification of spam emails, the value of C=0.5 is the best since it classified the least
##nonspam emails as spam.

finalmodel=ksvm(type~., data=spam[index[1:3501],], kernel="rbfdot", kpar=list(sigma=0.05), C=1)
finalpred=predict(finalmodel, newdata=test)
finalconfusion=table(test$type, finalpred)
finalmisclass=missclass(finalconfusion, test)
print(finalconfusion)
print(finalmisclass)

##Answer: The purpose of the parameter C is to put a weight to the cost function. The higher C the more cost will a
##constraint violation yield.

#Final model

finalmodel=ksvm(type~., data=spam, kernel="rbfdot", kpar=list(sigma=0.05), C=1)
```

Assignment 3 - NEURAL NETWORKS

```
library(neuralnet)
set.seed(1234567890)
Var <- runif(50, 0, 10)
trva <- data.frame(Var, Sin=sin(Var))
tr <- trva[1:25,] # Training
va <- trva[26:50,] # Validation
# Random initialization of the weights in the interval [-1, 1]
#set.seed(12345). Did not use this becaus it was not used in the code skeleton.
```

```
winit <- runif(31, -1, 1)
n = 10
SE_tr = vector("numeric", length = n)
SE_va = vector("numeric", length = n)
for(i in 1:n) {
  nn <- neuralnet(Sin ~ Var, data=tr, hidden = c(10), startweights = winit, threshold = i/1000 )

  p_tr = predict(nn, newdata = tr)
  SE_tr[i] = sum((tr$Sin - p_tr)^2)
  p_va = predict(nn, newdata = va)
  SE_va[i] = sum((va$Sin - p_va)^2)
}

which.min(SE_va) # 4/1000 has the lowest error.

plot(SE_tr, col = "red", ylim = c(0.001, 0.035), ylab = "Sum of Squared Error")
par(new=TRUE)
plot(SE_va, col = "blue", ylim = c(0.001, 0.035), ylab = "Sum of Squared Error")

#4/1000has the lowest Squared Error. Therefore 4/1000 is shosen as threshold.
plot(nn <- neuralnet(Sin ~ Var, data=tr, hidden = c(10), startweights = winit, threshold = 4/1000))

# Plot of the predictions (black dots) and the data (red dots)
plot(prediction(nn)$rep1)
points(trva, col = "red")
```

The Ultimate R Cheat Sheet – Data Management (Version 4)

Google "R Cheat Sheet" for alternatives. The best cheat sheets are those that you make yourself! Arbitrary variable and table names that are not part of the R function itself are highlighted in bold.

Import, export, and quick checks

- **dat1**=read.csv(**"name.csv"**) to import a standard CSV file (first row are variable names).
- attach(**dat1**) to set a table as default to look for variables. Use detach() to release.
- **dat1**=read.delim(**"name.txt"**) to import a standard tab-delimited file.
- **dat1**=read.fwf(**"name.prn"**, widths=c(**8,8,8**)) fixed width (3 variables, 8 characters wide).
- ?read.table to find out more options for importing non-standard data files.
- **dat1**=read.dbf(**"name.dbf"**) requires installation of the foreign package to import DBF files.
- head(**dat1**) to check the first few rows and variable names of the data table you imported.
- names(**dat1**) to list variable names in quotation marks (handy for copy and paste to code).
- data.frame(names(**dat1**)) gives you a list of your variables with the column number indicated, which can be handy for sub-setting a data table (see next page)
- nrow(**dat1**) and ncol(**dat1**) returns the number of rows and columns of a data table.
- length(**dat1**$**VAR1**[!is.na(**dat1**$**VAR1**)] returns a count of non-missing values in a variable.
- str(**dat1**) to check variable types, which is useful to see if the import executed correctly.
- write.csv(**results, "myresults.csv",** na="", row.names=F) to export data. Without the option statements, missing values will be represented by NA and row numbers will be written out.

Data types and basic data table manipulations

- There are three important variable types: numeric, character and factor (a double variable with a numeric and character value). You can query or assign types: is.factor() or as.factor().
- If you import a data table, variables that contain one or more character entries will be set to factor. You can force them to numeric with this: as.numeric(as.character(**dat1**$**VAR1**))
- After subsetting or modification, you might want to refresh factor levels with droplevels(**dat1**)
- Data tables can be set as.data.frame(), as.matrix(), as.distance()
- names(**dat1**)=c(**"ID", "X", "Y", "Z"**) renames variables. Note that the length of the vector must match the number of variable you have (four in this case).
- row.names(**dat1**)=**dat1**$**ID**. assigns an ID field to row names. Note that the default row names are consecutive numbers. In order for this to work, each value in the ID field must be unique.
- To generate unique and descriptive row names that may serve as IDs, you can combine two or more variables: row.names(**dat1**)=paste(**dat1**$**SITE, dat1**$**PLOT,** sep="-")
- If you only have numerical values in your data table, you can transpose it (switch rows and columns): **dat1_t**=t(**dat1**). Row names become variables, so run the row.names() function above first.
- **dat1**[order(**X**),] orders rows by variable X. dat[order(**X,Y**),] orders rows by variable X, then variable Y. **dat1**[order(**X,-Y**),]. Orders rows by variable X, then descending by variable Y.
- fix(**dat1**) to open the entire data table as a spreadsheet and edit cells with a double-click.

Creating systematic data and data tables

- c(1:10) is a generic concatenate function to create a vector, here numbers from 1 to 10.
- seq(0, 100, 10) generates a sequence from 0 to 100 in steps of 10.
- rep(5,10) replicates 5, 10 times. rep(c(1,2,3),2) gives 1 2 3 1 2 3. rep(c(1,2,3), each=2) gives 1 1 2 2 3 3. This can be useful to create data entry sheets for experimental designs.
- data.frame(**VAR1**=c(1:10), **VAR2**=seq(10, 100, 10), **VAR3**=rep(c(**"this", "that"**),5)) creates a data frame from a number of vectors.
- expand.grid(**SITE**=c(**"A","B"**),**TREAT**=c(**"low","med","high"**), REP=c(1:5)) is an elegant method to create systematic data tables.

Creating random data and random sampling

- `rnorm(10)` takes 10 random samples from a normal distribution with a mean of zero and a standard deviation of 1
- `runif(10)` takes 10 random samples from a uniform distribution between zero and one.
- `round(rnorm(10)*3+15))` takes 10 random samples from a normal distribution with a mean of 15 and a standard deviation of 3, and with decimals removed by the rounding function.
- `round(runif(10)*5+15)` returns random integers between 15 and 20, uniformly distributed.
- `sample(c("A","B","C"), 10, replace=TRUE)` returns a random sample from any custom vector or variable with replacement.
- **sample1**=**dat1**`[sample(1:nrow(`**dat1**`), 50, replace=FALSE),]` takes 50 random rows from dat1 (without duplicate sampling). This can be handy for bootstrapping or to run quick test analyses on subsets of very large datasets.

Sub-setting data tables, conditional subsets

- **dat1**`[1:10, 1:5]` returns the first 10 rows and the first 5 columns of table dat1.
- **dat2**=**dat1**`[50:70,]` returns a subset of rows 50 to 70.
- **cleandata**=**dat1**`[-c(2,7,15),]` removes rows 2, 7 and 15.
- **selectvars**=**dat1**`[,c("ID","YIELD")]` sub-sets the variables ID and YIELD
- **selectrows**=**dat1**`[`**dat1**`$`**VAR1**`=="Site 1",]` sub-sets entries that were measured at Site 1. Possible conditional operators are == equal, != non-equal, > larger, < smaller, >= larger or equal, <= smaller or equal, & AND, | OR, ! NOT, () brackets to order complex conditional statements.
- **selecttreats**=**dat1**`[`**dat1**`$`**TREAT** `%in% c("CTRL", "N", "P", "K"),]` can replace multiple conditional == statements linked together by OR.

Transforming variables in data tables, conditional transformations

- **dat2**=`transform(`**dat1, VAR1**=**VAR1**`*0.4)`. Multiplies VAR1 by 0.4
- **dat2**=`transform(`**dat1, VAR2**=**VAR1**`*2)`. Creates variable VAR2 that is twice the value of VAR1
- **dat2**=`transform(`**dat1, VAR1**=`ifelse(`**VAR3**`=="Site 1", VAR1`*`0.4, VAR1`)) Multiplies VAR1 by 0.1 for entries measured at Site 1. For other sites the value stays the same. The general format is ifelse(condition, value if true, value if false).
- The `vegan` package offers many useful standard transformations for variables or an entire data table: **dat2**=`decostand(`**dat1,**`"standardize")` Check out `?decostand` to see all transformations.

Merging data tables

- **dat3**=`merge(`**dat1,dat2,**`by="ID")` merge two tables by ID field.
- **dat3**=`merge(`**dat1,dat2,**`by.x="ID",by.y="STN")` merge by an ID field that is differently named in the two datasets.
- **dat3**=`merge(`**dat1,dat2,**`by=c(`**"LAT","LONG"**`))` merge by multiple ID fields.
- **dat3**=`merge(`**dat1,dat2,**`by.x=`**"ID"**`,by.y=`**"ID"**`,all.x=T,all.y=F)` left merge; a`ll.x=F, all.y=T` right merge; `all.x=T,all.y=T` keep all rows; `all.x=F,all.y=F` keep matching rows.
- `cbind(`**dat1,dat2**`)` On very rare occasions, you merge data without a criteria (ID). This is generally dangerous, because the commands will slap the two tables together without checking the order!
- **dat3**=`rbind(`**dat1,dat2**`)` adding rows of two data tables. The variables have to match exactly and you will get error messages if they don't match. So, unlike cbind(), rbind() is generally safe to use.
- **dat3**=`rbind.fill(`**dat1,dat2**`)` will force non-matching datasets together, filling missing values and executing variable type conversions where appropriate. Requires the `reshape` package.

Summary statistics for variables and tables

- `mean() weighted.mean(,) median() max() min() range() which.max() which.min() var() sd() quantile() quantile(,c(0.025,0.05,0.95,0.975)) rank(x)` some descriptive statistical functions for variables or vectors. For all functions, and

important option is `na.rm=T`, which means that missing values are ignored in the calculations, e.g. `mean(VAR1, na.rm=T)`. Without that option, the function returns missing values as a result of missing values in the input.

- `rowSums()`, `colSums()`, `rowMeans()` or `colMeans()` applies functions to rows or columns of a table. For example, `rowsum(dat1[,10:15])` will return the row-sums of the variables in columns 10 to 15. Don't forget `na.rm=T`.
- `apply(dat,1,max)` apply any function (e.g. **max**), to either rows (**1**) or columns (**2**) of a table (**dat**).

Pivot table functionality

- The functions aggregate and ddply can be used to summarize data similarly to working with Excel pivot tables. Aggregate has simpler syntax if you have many variables that you want to summarize in the same way; ddply is better if you have few variables but want several custom summary statistics.
- `aggregate(dat1[,4:9], by=list(TREAT1=dat1$TREAT1, TREAT2=dat2$TREAT2), mean)` calculates the means of a number of numerical variables in columns 4 to 9 for two treatments.
- `ddply(dat1,.(TREAT), summarise, mVAR1=mean(VAR1))` returns means of VAR1 by a class variables TREAT. This requires installation of the library `plyr`.
- `ddply(dat1,.(TREAT1, TREAT2), summarise, cVAR1=length(VAR1[!is.na(VAR1)])` returns a count of non-missing values in variable VAR1 for each combination of two class variables.
- `ddply(dat1,.(TREAT1, TREAT2), summarise, mVAR1=mean(VAR1, na.rm=T), seVAR1=sd(VAR1, na.rm=T)/sqrt(length(VAR1[!is.na(VAR1)]))).` A clever piece of code to calculate means and standard errors, with missing values being properly handled.

Long-to-wide and wide-to-long data table conversions

- First we generate an artificial dataset to play with (copy and paste to R to see what it does):
  ```
  long=expand.grid(SITE=c("A","B"),TREAT=c("low","med","high"), REP=c(1:5))
  long$YIELD=round(rnorm(10)*5+15); long
  ```
- Long-to-wide conversion with the `reshape2` package, where SITE and REP remain columns, but the treatment levels of TREAT now become several new columns:
  ```
  wide=dcast(long, SITE+REP~TREAT, value.var="YIELD")
  ```
  Wide-to-long conversion back to what we had before. The variables that you want to maintain as columns are SITE and REP, all others will be gathered into a new variable where the remaining columns become treatment levels. You have to fix the variable names to get to the original long:
  ```
  long2=melt(wide, id.vars=c("SITE","REP"))
  names(long2)=c("SITE","REP","TREAT","YIELD")
  ```

Dealing with missing values

- `transform(dat1, VAR1=ifelse(is.na(VAR1),0,VAR1))` sets missing values in variable VAR1 to 0. You may do this if missing has a biological meaning of zero, e.g. zero productivity.
- `transform(dat1, VAR1=ifelse(VAR1==0,NA,VAR1))` … or vice versa if zero really means that the measurement was not taken.
- `dat1[is.na(dat1)]=0` sets missing values to 0 in entire dataset.
- `dat1[dat1==0]=NA` ... vice versa.
- `dat2=na.omit(dat1)` delete rows with missing values in any variable
- `dat2=dat1[!is.na(dat1$VAR1),]` delete rows with missing values in VAR1
- `dat2=transform(dat1, VAR2=ifelse(is.na(VAR1),NA,VAR2))` modify a second variable (here: set to missing) based on missing values in a first variable.

Dealing with duplicate data entries or IDs:

- `unique(dat1)` or `dat1[!duplicated(dat1),]` removes exact duplicate rows.
- `dat1[duplicated(dat1),]` returns the duplicate rows.
- `dat1[!duplicated(dat1[,c("ID")]),]` removes all rows with duplicate IDs (first is kept).
- `dat1[duplicated(dat1[,c("ID")]),]` returns the rows with duplicate IDs.

Loops and automation

- `v1=vector(length=20)` initializes an empty vector with 20 elements. This is often required as an initial statement to subsequently write results of a loop into an output vector or output table.
- `m1=matrix(nrow=20, ncol=10)` similarly initializes an empty matrix with 20 rows and 10 columns.
- `for (i in 1:10) {` one or more operations with `v1[i]` or `m1[,i]` `}`
- `for (i in 1:10) { for (j in 1:20) {` one or more operations with `m1[j,i]` `}}`
- Example for an application, where a for-loop is used to calculate cumulative values. Copy and paste the code below into R to see what it does.

```
dat=round(rnorm(10)+2)
cum=vector(length=10)
cum[1]=dat[1]
for (i in 2:length(dat)) { cum[i]=cum[i-1]+dat[i] }
cbind(dat,cum)
```

- Example for an application where a for-loop allows automatic data processing of multiple files in a directory. This batch-converts DBF format files into CSV format files. With similar code, you could merge a large number of files into one master file, or do manipulations or analysis on multiple files consecutively.

```
library(foreign)
setwd("C:/your path/")
a<-list.files(); a
for (name in a) { dat1=read.dbf(name)
  write.csv(dat1, paste(name,".csv"), row.names=F, quote=F) }
```

Handy built-in functions

- `paste("hello", "world")` joins vectors after converting them to characters. The `sep=""` option can place any character string or nothing between values (a single space is the default)
- `substr("Year 1998",6,9)` extracts characters from start to stop position from vector
- `tolower("Year 1998")` convert to lowercase - handy to correct inconsistencies in data entry.
- `toupper("Year 1998")` convert to uppercase
- `nchar("Year 1998")` number of characters in a string, allows you to substring the last four digits of a variable regardless of length, for example: `substr(VAR1,nchar(VAR1)-3,nchar(VAR1))`
- Plenty of math functions, of course: `log(VAR1), log10(VAR1), log(VAR1,2), exp(VAR1), sqrt(VAR1), abs(VAR1), round(VAR1,2)`

Programming custom functions

- You can program your own functions, if something is missing, or if you want to utilize a bunch of code over and over to make similar calculations. Here is a clever example for calculating the statistical mode of a variable, which is missing from the built-in R functions.

```
mode=function(input){ freq=table(as.vector(input))
                      descending_freq=sort(-freq)
                      mode=names(descending_freq)[1]
                      as.numeric(mode)
                      }
VAR1=c(1,3,3,2,3,2,2,3,5,3)
mode(VAR1)
```

More information, help, and on-line resources

- Adding a question mark before a command or functions brings up a help file. E.g. `?paste`. Be sure to check out the example code at the end of the help file, which often helps to understand the syntax.
- More information and R resources can be found with the search engine http://www.rseek.org