

Types of data sets

- **Training data** (training set D):

used for fitting the model

- Supervised learning: w_i in $P(y|x, w_1, \dots w_k)$ 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 x

x	y
1.3	?
2.9	?

732A99/TDDE01

Logistic regression

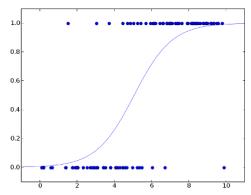
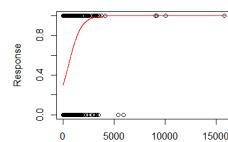
- Data $Y_i \in \{\text{Spam}, \text{Not Spam}\}, X_i = \#\text{of a word}$

- Model: $p(Y = \text{Spam}|w, x) = \frac{1}{1+e^{-w_0-w_1x}}$

- Fitting: maximum likelihood

- Prediction : $p(\text{spam}) = p(Y = \text{spam}|x)$

We can also make point predictions
-how?



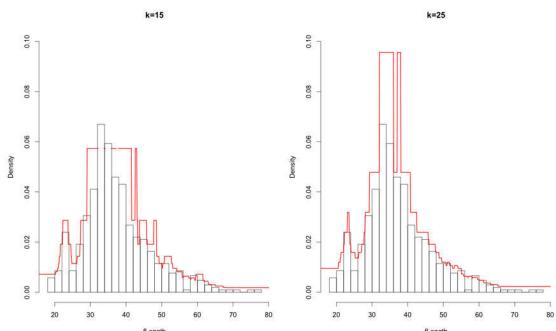
732A99/TDDE01

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 $p(x|K)$

732A99/TDDE01

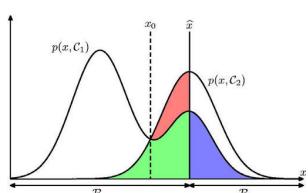
K-nearest neighbor density estimation



732A99/TDDE01

K-nearest neighbor density estimation

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



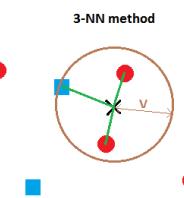
732A99/TDDE01

K-nearest neighbor classification

- Given N observations (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 $Y = C_i$



732A99/TDDE01

Bayesian classification

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

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

$$p(C_i | \mathbf{x}) = \frac{p(\mathbf{x} | C_i) p(C_i)}{p(\mathbf{x})}$$
- We get

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

732A99/TDDE01

K-nearest neighbor classification

Algorithm

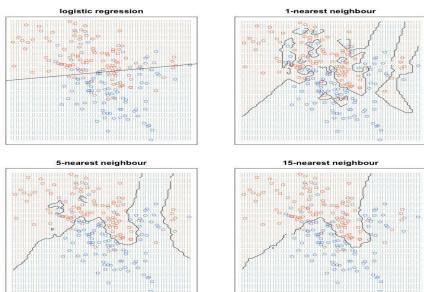
- Given training set D , number K , and test set T
- For each $\mathbf{x} \in T$
 - For each $i = 1, \dots, M$
 - $p'(C_i | \mathbf{x}) = \frac{K_i}{K}$
 - Compute $l = \arg \max_{i \in \{1, \dots, m\}} p'(C_i | \mathbf{x})$
 - Predict $\hat{Y}(\mathbf{x}) = C_l$

Majority voting: prediction for \mathbf{x} is defined by majority voting of K neighbors

732A99/TDDE01

K-nearest neighbor example

Why classification results are so different for K-NN?

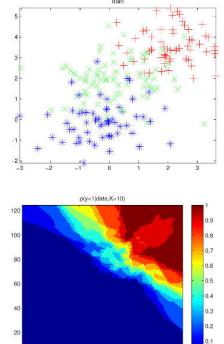


732A99/TDDE01

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



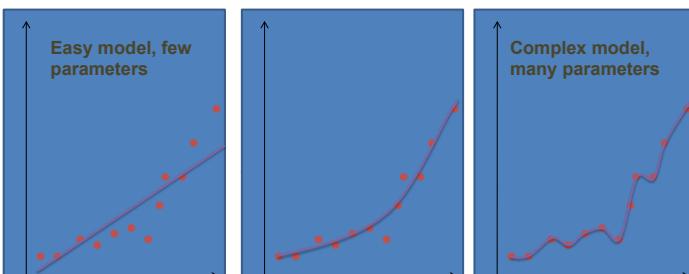
Nonparametric models

- Number of parameters (complexity) grows with training data
- Example: K-NN classifier

732A99/TDDE01

Overfitting

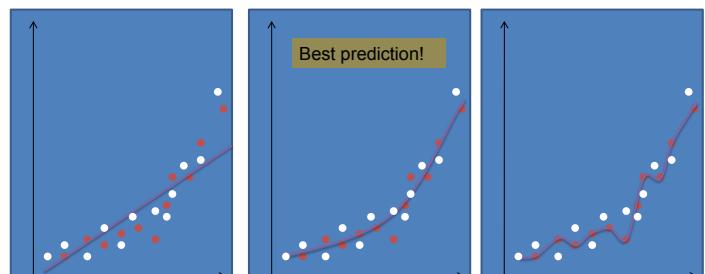
- Which model feels appropriate?



732A99/TDDE01

Overfitting

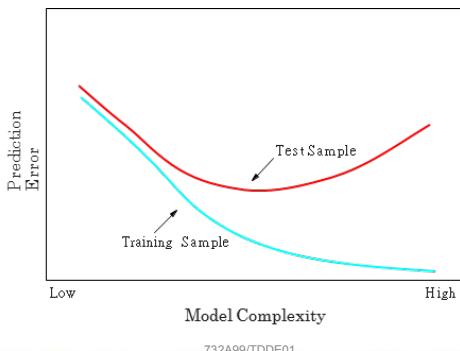
Now new data from the same process



732A99/TDDE01

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

732A99/TDDE01

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**

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

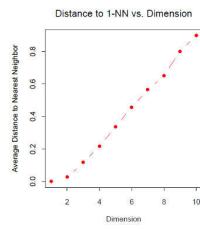
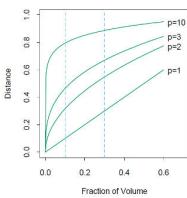
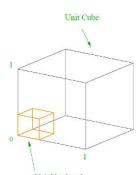
732A99/TDDE01

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 misleading
 - Distance-based methods affected

732A99/TDDE01

Curse of dimensionality



732A99/TDDE01

Curse of dimensionality

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

732A99/TDDE01

Basics of Statistics

Lecture 1b

Probability

How likely it is that some event will happen?

Idea:

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

732A99/TDDE01

2

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

732A99/TDDE01

3

Properties and definitions

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

732A99/TDDE01

4

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?

732A99/TDDE01

5

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 occurred
- New (updated) knowledge about B
 - Posterior probability $P(B|A)$

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

732A99/TDDE01

6

Random variables

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

Examples:

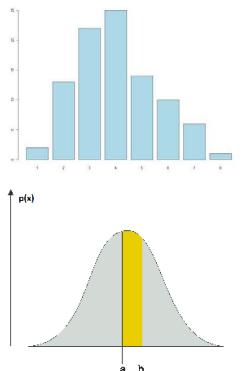
- $X = \{\text{amount of times the word "crisis" can be found in financial documents}\}$
 - $P(X=3)$
- $X = \{\text{Time to download a specific file to a specific computer}\}$
 - $P(X=0.36 \text{ min})$

732A99/TDDE01

7

Distributions

- Discrete
 - Probability mass function $P(x)$ for all feasible x
- Continuous
 - 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 $F(x) = \int_0^x p(t)dt$



732A99/TDDE01

8

Expected value and variance

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

732A99/TDDE01

9

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:

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

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

732A99/TDDE01

10

Bayes theorem

For random variables:

Bayes Theorem

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

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

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



732A99/TDDE01

11

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, winning a lottery,..

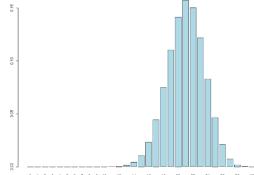
732A99/TDDE01

12

Some conventional distributions

Binomial distribution

- Sequence of n Bernoulli events
 - $X=\{\text{Amount of successes among these events}\}, X=0,\dots,n$
- $$P(X = r) = \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r}$$
- $EX = np$
 - $Var(X) = np(1-p)$



732A99/TDDE01

13

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 $\lambda=np$ is mean amount of persons that call between 13.00 and 14.00
- $X=\{\text{amount of persons that have called between 13.00 and 14.00}\}$

732A99/TDDE01

14

Poisson distribution

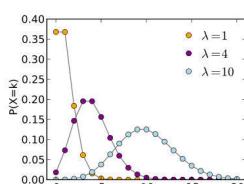
- $P(X = r) = \lim_{n \rightarrow \infty} \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r}$

It can be shown that

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

- $E(X) = \lambda$

- $Var(X) = \lambda$



732A99/TDDE01

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 $np \leq 10$

732A99/TDDE01

16

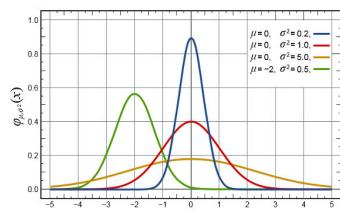
Normal distribution

Appears in almost all applications

- Difference between the times required to download two specific documents to a specific computer

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

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



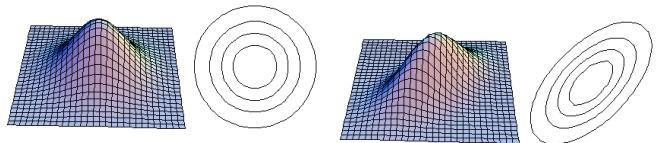
732A99/TDDE01

17

Multivariate distributions

Probability of two variables having certain values at the same time

- P.D.F. $p(x,y)$
- Correlation



732A99/TDDE01

18

Basic ML ingredients

- Data D : observations
 - Features X_1, \dots, X_p
 - Targets Y_1, \dots, Y_r
- 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
- Predict new data X^{new} by using the fitted model

Case	X_1	X_2	Y
1			
2			
...			

732A99/TDDE01

19

Probabilistic models

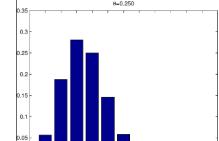
- A distribution $p(x|w)$ or $p(y|x, w)$

- Example:

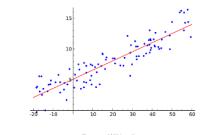
- $x \sim Bin(n, \theta)$

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

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



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



Source: Wikipedia

732A99/TDDE01

20

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

732A99/TDDE01

21

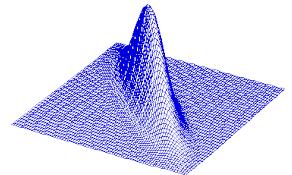
Fitting a model

- Frequentist principle: **Maximum likelihood** principle

- Compute likelihood $p(D|w)$

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

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



- Maximize the likelihood and find the optimal w^*

732A99/TDDE01

22

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...

732A99/TDDE01

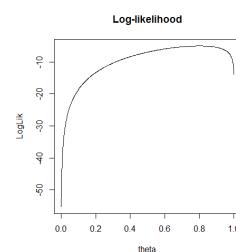
23

Fitting a model

Example: tossing a coin.

$$D = \{0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1\},$$

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



)E01

24

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

732A99/TDDE01

25

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^*

732A99/TDDE01

26

Fitting a model

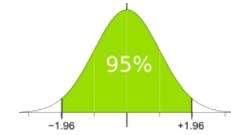
- How to chose the prior?
 - Expert knowledge about the phenomenon
 - Forcing a model to have a certain structure
 - Example: decision trees: prior prefers smaller trees
http://en.wikipedia.org/wiki/Conjugate_prior
 - Conjugacy
 - Distribution of the posterior is the same type as the distribution of the likelihood or prior
- Prior is the most controversial about Bayesian methods, but
 - When $N \rightarrow \infty$, data overwhelms the prior

732A99/TDDE01

27

Measuring uncertainty

- Confidence interval** (frequentist)
 - Model $p(x|w)$ is known
 - \hat{w} is a function of x by ML
 - Derive distribution of \hat{w}
 - Compute quantiles
- Credible interval** (Bayes)
- Prediction interval** (models)



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

732A99/TDDE01

28

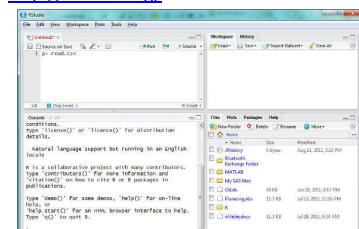
Introduction to R

Lecture 1c

732A99/TDDE01

R

- R is a **programming language** of a higher-level
- Constantly increasing amount of packages (new research)
- Free of charge
- Website: <http://www.r-project.org/>
- Code Editor: <http://rstudio.org/>



732A99/TDDE01

Overview

- Linear regression
- Ridge Regression
- Lasso
- Variable selection

732A99/TDDE01

2

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

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

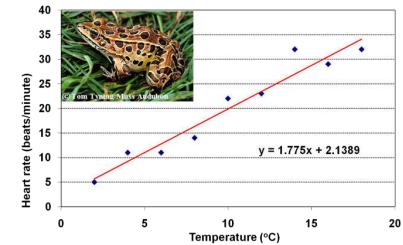
Terminology:

w_0 : intercept (or bias)

w_1 : regression coefficient

Response

The target responds directly and linearly to changes in the feature



732A99/TDDE01

3

Ordinary least squares regression (OLS)

Model:

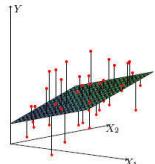
$$y \sim N(\mathbf{w}^T \mathbf{x}, \sigma^2)$$

where

$$\mathbf{w} = \{w_0, \dots, w_d\}$$

$$\mathbf{x} = \{1, x_1, \dots, x_d\}$$

Why is "1" here?



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

732A99/TDDE01

4

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

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

Is equivalent to minimizing

$$RSS(\mathbf{w}) = \sum_{i=1}^n (Y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

732A99/TDDE01

5

Matrix formulation of OLS regression

Optimality condition:

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

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

732A99/TDDE01

6

Parameter estimates and predictions

- Least squares estimates of the parameters

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

- Predicted values

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



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

Why is it called so?

732A99/TDDE01

7

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 $\hat{Y} = S(X)Y$

$$df = \text{trace}(S)$$

- For linear regression, degrees of freedom is

$$df = \text{trace}(P) = p$$

732A99/TDDE01

8

Different types of features

- Interval variables

- Numerically coded ordinal variables

• (small=1, medium=2, large=3)

- Dummy coded qualitative variables

Example of dummy coding:

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

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:

$$\phi_1(x_1) = x_1$$

$$\phi_2(x_1) = x_1^2$$

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

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

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

732A99/TDDE01

9

Basis function expansion

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

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

732A99/TDDE01

10

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
```

732A99/TDDE01

11

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)

> 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  0.288 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:  0.8997, Adjusted R-squared:  0.8942 
F-statistic: 86.26 on 1 and 57 DF,  p-value: < 2.2e-16
```

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

Inputs:
 X_1 = Manufacturing year
 X_2 = Milage (km)
 X_3 = Equipment (0 or 1)

732A99/TDDE01

12

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
```

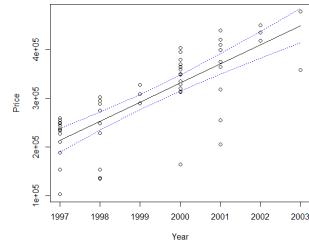
732A99/TDDE01

13

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)
```



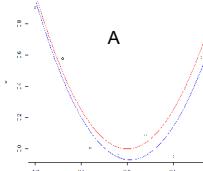
732A99/TDDE01

14

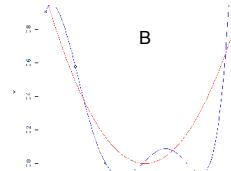
Ridge regression

- Problem: linear regression can overfit:

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



A



B

732A99/TDDE01

15

Ridge regression

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

$$\text{minimize } -\log \text{likelihood} + \lambda_0 \|w\|_2^2$$

→ **L_2 regularization**

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

$$\hat{w}^{\text{ridge}} = \underset{w}{\operatorname{argmin}} \left\{ \sum_{i=1}^N (y_i - w_0 - w_1 x_{1j} - \dots - w_p x_{pj})^2 + \lambda \sum_{j=1}^p w_j^2 \right\}$$

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

Equivalent form

$$\begin{aligned} \hat{w}^{\text{ridge}} &= \underset{w}{\operatorname{argmin}} \sum_{i=1}^N (y_i - w_0 - w_1 x_{1j} - \dots - w_p x_{pj})^2 \\ &\text{subject to } \sum_{j=1}^p w_j^2 \leq s \end{aligned}$$

Solution

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

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

How do we compute degrees of freedom here?

Hat matrix

732A99/TDDE01

16

732A99/TDDE01

17

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 orthogonal (not realistic), $X^T X = I \rightarrow \hat{w}^{\text{ridge}} = \frac{1}{1+\lambda} w^{\text{linreg}}$ → coefficients are equally shrunk
- Ridge regression is particularly useful if the explanatory variables are strongly correlated to each other.**
 - Correlated variables often correspond large $w \rightarrow$ shrunk
- Degrees of freedom decrease when λ increases
 - $\lambda = 0 \rightarrow d.f. = p$

Properties

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

732A99/TDDE01

18

732A99/TDDE01

19

Ridge regression

- Bayesian view

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

$$y \sim N(y|w_0 + Xw, \sigma^2 I)$$

$$w \sim N\left(0, \frac{\sigma^2}{\lambda} I\right)$$

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

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

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

732A99/TDDE01

20

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

- Cycle time
- Memory
- Channels
- ...

Build model predicting performance



732A99/TDDE01

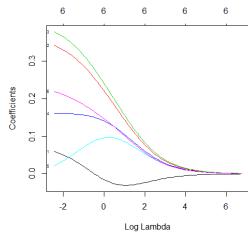
21

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)
```



732A99/TDDE01

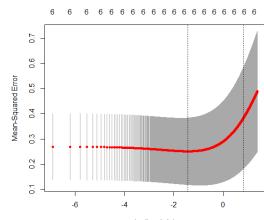
22

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 "dgCMatrix"
 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

23

Ridge regression

- How good is this model in prediction?

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

covariates=train[,1:6]
response=train[, 7]
model=cv.glmnet(as.matrix(covariates), response, alpha=1,family="gaussian",
lambda=seq(0.1,0,0.001))
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)
Note that data are so small so numbers
change much for other train/test
```

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

732A99/TDDE01

24

LASSO

- **Idea:** Similar idea to Ridge
- Minimize minus loglikelihood plus **linear** penalty factor $\rightarrow \mathbf{I}_1$ regularization
- Given that model is Gaussian, we get **LASSO** (least absolute shrinkage and selection operator):

$$\hat{w}^{LASSO} = \operatorname{argmin}_{\mathbf{w}} \left\{ \sum_{i=1}^N (y_i - w_0 - w_1 x_{1j} - \dots - w_p x_{pj})^2 + \lambda \sum_{j=1}^p |w_j| \right\}$$

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



732A99/TDDE01

25

LASSO

- Equivalently

$$\hat{w}^{lasso} = \underset{\sum_{j=1}^p |w_j| \leq s}{\operatorname{argmin}} \sum_{i=1}^N (y_i - w_0 - w_1 x_{1j} - \dots - w_p x_{pj})^2$$

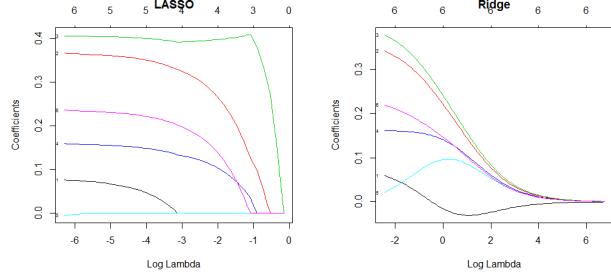
732A99/TDDE01

26

LASSO vs Ridge

- LASSO yields sparse solutions!

Example Computer hardware data



732A99/TDDE01

27

LASSO vs Ridge

- Only 5 variables selected by LASSO

```
> coef(model1, 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            2.287134e-01
v8            2.287134e-01
> sum((ynew-mean(y))^2)/sum((y-mean(y))^2)
[1] 0.5826904
> sum((ynew-y)^2)
[1] 16.63756
```

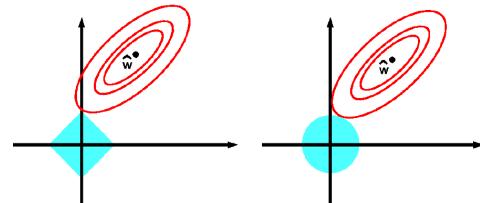
732A99/TDDE01

28

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)



732A99/TDDE01

29

LASSO properties

- 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,

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

- No explicit formula for \hat{w}^{lasso}

- Optimization algorithms used

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

732A99/TDDE01

30

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

732A99/TDDE01

31

Variable selection

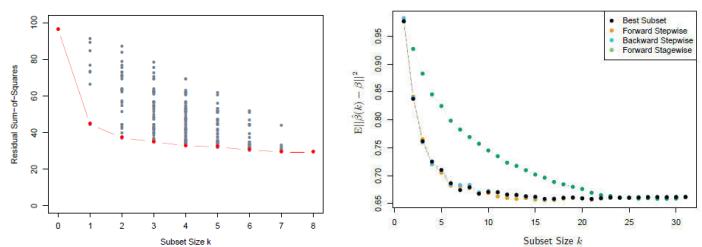
Alternative 1: Variable subset selection

- 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")

732A99/TDDE01

32

RSS and MSE depend on k



732A99/TDDE01

33

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.fr)

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-15  2.827e-07  2.033  0.0054 ** 
V3           3.661e-01  4.312e-02  8.490  4.34e-15 ***
V4           4.056e-01  4.684e-02  8.895  1.18e-15 ***
V5           1.593e-01  2.042e-02  7.766  6.0e-06 ***
V6           2.360e-01  3.356e-02  7.031  3.40e-11 *** 
V8           9.7492   37.852   -345.11 
V9           10.4837   38.586   -341.09 
              DF Sum of Sq   RSS   AIC  
<none>             28.117 -407.25 
- V7          1   0.0139  28.103 -405.35 
- V3          1   1.0819  29.185 -399.46 
- V6          1   2.3150  31.160 -387.57 
- V8          1   9.7492  37.852 -345.11 
- V5          1   10.4837 38.586 -341.09 

```

732A99/TDDE01

34

Model selection

Lecture 1e

Overview

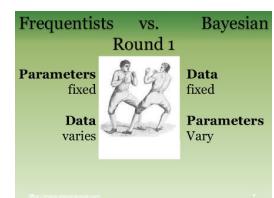
- Model fitting
- Model selection

732A99/TDDE01

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?

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

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

2

732A99/TDDE01

3

An estimator

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

Example: Linear regression

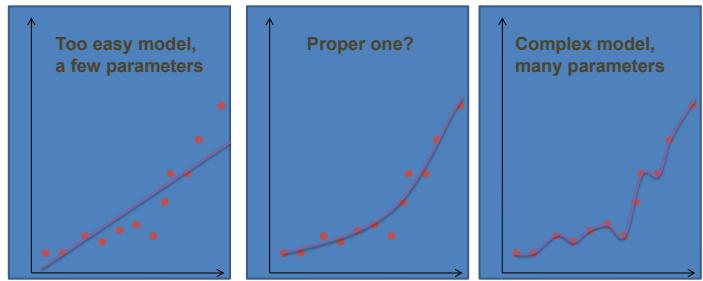
- Estimator 1: $w = (X^T X)^{-1} X^T Y$ (maximum likelihood)
- Estimator 2: $w = (0, \dots, 0, 1)$
- Which one is better?
 - A comparison strategy is needed!

732A99/TDDE01

4

Overfitting

- Complex model can overfit your data



732A99/TDDE01

5

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

732A99/TDDE01

6

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}

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

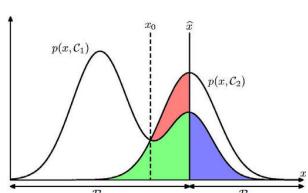
732A99/TDDE01

7

Model selection

Example: Spam classification

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



732A99/TDDE01

8

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:

$$L(Y, \hat{y}) = \begin{cases} 10 + \frac{\hat{y}}{Y}, & \hat{y} \geq Y \\ 1000, & \hat{y} < Y \end{cases}$$

Example: Compute loss function related to

- Normal distribution

Guess why such loss function was chosen

732A99/TDDE01

9

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

732A99/TDDE01

10

Model selection

- Problem:** true model and true 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**)

732A99/TDDE01

11

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**)

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

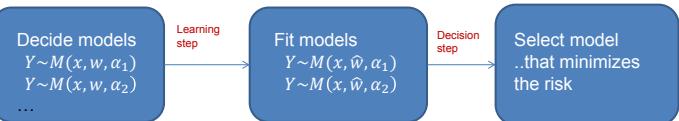
- Estimator is fit by Maximum Likelihood using training set
- Risk estimated by using validation set
- Model with minimum empirical risk is selected

732A99/TDDE01

12

General model selection strategy

- Given data $D = \{X_i, Y_i, i = 1 \dots n\}$



- When fitting data, Maximum Likelihood is usually used

- α_i can be different things:
 - Type of distribution
 - Number of variables in the model
 - Regularization parameter value
 - ...

732A99/TDDE01

13

Holdout method

Divide into training, validation and test sets



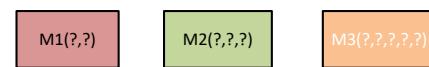
- Choose proportions in some way

732A99/TDDE01

14

Holdout method

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

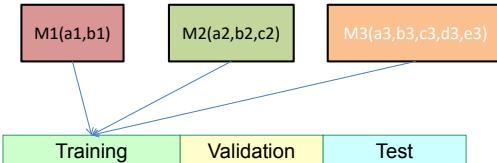


732A99/TDDE01

15

Holdout method

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

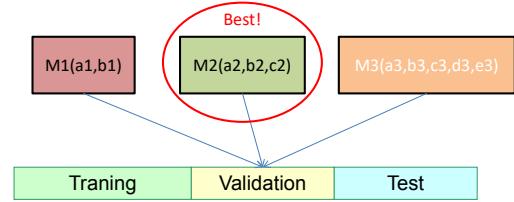


732A99/TDDE01

16

Holdout method

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

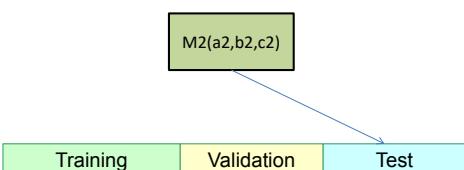


732A99/TDDE01

17

Holdout method

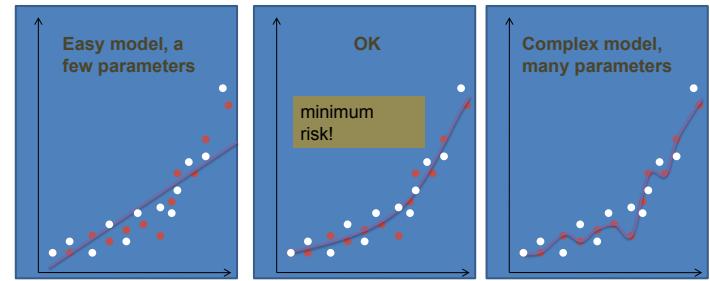
- Test set is used to test a performance on a new data



732A99/TDDE01

18

Holdout method



732A99/TDDE01

19

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)
id1=sample(1:n, floor(n*0.4))
train=data[id1,]

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

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

732A99/TDDE01

20

Bias-variance tradeoff

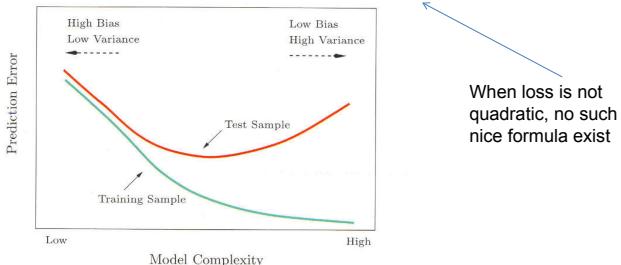
- Bias of an estimator $Bias(\hat{y}(x_0)) = E[\hat{y}(x_0) - f(x_0)]$, $f(x_0)$ is expected response
 - If $Bias(\hat{y}(x_0)) = 0$, the estimator is **unbiased**
 - ML estimators are asymptotically unbiased if the model is enough complex
 - However, unbiasedness does not mean a good choice!

732A99/TDDE01

21

Bias-variance tradeoff

- Assume loss is $L(Y, \hat{y}) = (Y - \hat{y})^2$
 $R(Y(x_0), \hat{y}(x_0)) = \sigma^2 + Bias^2(\hat{y}(x_0)) + Var(\hat{y}(x_0))$



732A99/TDDE01

22

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):

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

732A99/TDDE01

23

Cross-validation

Cross-validation



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

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

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?

732A99/TDDE01

24

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

732A99/TDDE01

25

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):

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

732A99/TDDE01

26

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

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

732A99/TDDE01

27

Model selection

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

- Cycle time
- Memory
- Channels
- ...

Build model predicting performance



732A99/TDDE01

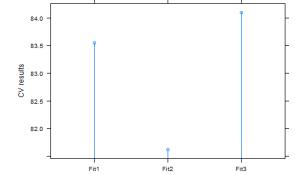
28

Cross-validation

- 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)
```



732A99/TDDE01

29

Linear classification methods

Lecture 2a

Overview

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

732A99/TDDE01

2

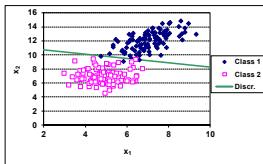
Classification

Classifiers

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

Classification problem:

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



732A99/TDDE01

3

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

Disadvantages 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 → ???
 - $P(\text{Spam}|A)=0.99, P(\text{Spam}|B)=0.45 \rightarrow$ better decision can be made

732A99/TDDE01

4

Bayesian decision theory

Loss matrix

- Machine learning models estimate $p(y|x)$ or $p(y|x, \hat{w})$
- Transform probability into action → which value to predict? → decision step
 - $p(Y = \text{Spam}|x) = 0.83 \rightarrow$ 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**

- 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

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

732A99/TDDE01

5

732A99/TDDE01

6

Loss and decision

- Expected loss minimization

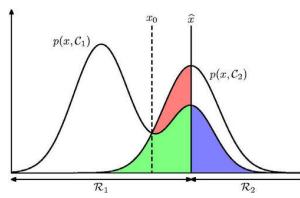
– R_j : classify to C_j

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

- Choose such R_j that EL is minimized**

- Two classes

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



732A99/TDDE01

7

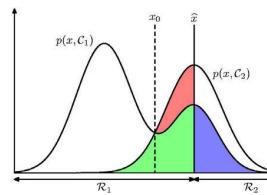
Loss and decision

- Loss minimization

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

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

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



732A99/TDDE01

8

Loss and decision

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

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

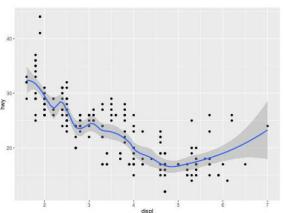
732A99/TDDE01

9

Loss and decision

- Continuous targets: squared loss

– Given a model $p(x, y)$,
minimize
 $EL = \int L(y, \hat{Y}(x)) p(x, y) dxdy$



- Using **square loss**, the optimal is posterior mean

$$\hat{Y}(x) = \int y p(y|x) dy$$

732A99/TDDE01

10

ROC curves

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

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

732A99/TDDE01

11

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**

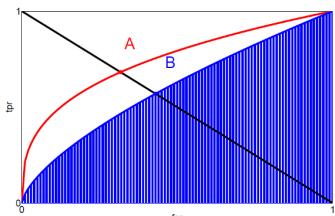
$$Precision = \frac{TP}{TP + FP}$$

732A99/TDDE01

12

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**)



732A99/TDDE01

13

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

$$p(Y = 1|w, x) = \frac{1}{1 + e^{-w^T x}}$$

732A99/TDDE01

14

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

732A99/TDDE01

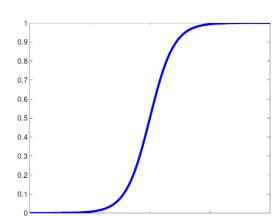
15

Logistic regression

- Discriminative model
- Model for binary output
 - $C = \{C_1 = 1, C_2 = 0\}$
 $p(Y = C_1|X) = \text{sigm}(w^T x)$
- Alternatively
 $Y \sim \text{Bernoulli}(\text{sigm}(a)), a = w^T x$
 $\text{sigm}(a) = \frac{1}{1 + e^{-a}}$

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

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



732A99/TDDE01

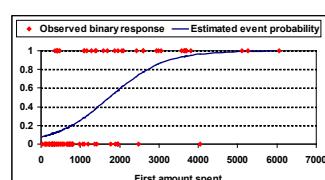
16

Logistic regression

- Logistic model- yet another form
 $\ln \frac{p(Y = 1|X = x)}{P(Y = 0|X = x)} = \ln \frac{p(Y = 1|X = x)}{1 - P(Y = 1|X = x)} = \text{logit}(p(Y = 1|X = x)) = w^T x$
- Here $\text{logit}(t) = \ln \left(\frac{t}{1-t} \right)$
- Note $p(Y|X)$ is connected to $w^T x$ via logit link

The log of the odds is linear in x

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

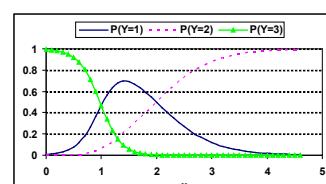


732A99/TDDE01

17

Logistic regression

- When Y is categorical,
 $p(Y = C_i|x) = \frac{e^{w_i^T x}}{\sum_{j=1}^K e^{w_j^T x}} = \text{softmax}(w_i^T x)$
- Alternatively
 $Y \sim \text{Multinoulli}(\text{softmax}(w_1^T x), \dots, \text{softmax}(w_K^T x))$



732A99/TDDE01

18

Logistic regression

Fitting logistic regression

- In binary case,

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

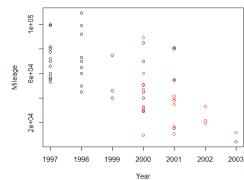
732A99/TDDE01

Logistic regression

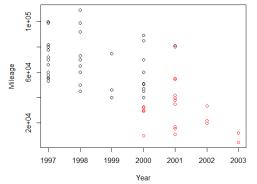
- In R, use `glm()` with family="binomial"
 - Predicted probabilities: `predict(fit,newdata, type="response")`

Example `Equipment=f(Year, mileage)`

Original data



Classified data



732A99/TDDE01

20

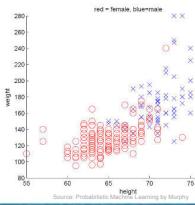
Quadratic discriminant analysis

- Generative classifier
- Main assumptions:

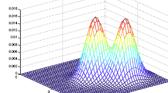
x is now random as well as y

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

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



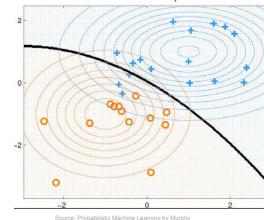
Source: Probabilistic Machine Learning by Murphy



Quadratic discriminant analysis

- If parameters are estimated, classify:

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



Source: Probabilistic Machine Learning by Murphy

732A99/TDDE01

22

Linear discriminant analysis (LDA)

- Assumption $\Sigma_i = \Sigma, i = 1, \dots, K$
- Then $p(y = c_i|x) = \text{softmax}(w_i^T x + w_{0i}) \rightarrow$ exactly the same form as the logistic regression

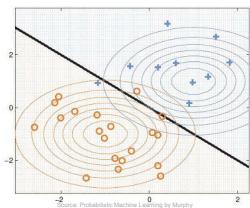
$$\begin{aligned} w_{0i} &= -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log \pi_i \\ w_i &= \Sigma^{-1} \mu_i \end{aligned}$$

- Decision boundaries are linear

Discriminant function:

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

732A99/TDDE01



23

Linear discriminant analysis (LDA)

- Difference LDA vs logistic regression??

Coefficients will be estimated differently! (models are different)

- How to estimate coefficients

find MLE.

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

Sample mean and sample covariance are MLE!

If class priors are parameters (**proportional priors**),

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

732A99/TDDE01

24

LDA and QDA: code

- Syntax in R, library MASS

```
lda(formula, data, ..., subset, na.action)
```

- Prior – class probabilities
- Subset – indices, if training data should be used

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

```
predict(..)
```

732A99/TDDE01

25

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
```

732A99/TDDE01

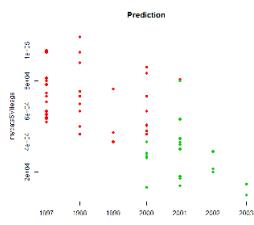
26

LDA: output

- Misclassified items

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

```
> table(Pred$class, mydata$Equipment)
   0   1
0 31  6
1  7 15
```

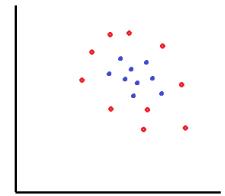


732A99/TDDE01

27

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



732A99/TDDE01

28

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 $\Sigma \rightarrow$ 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.

732A99/TDDE01

29

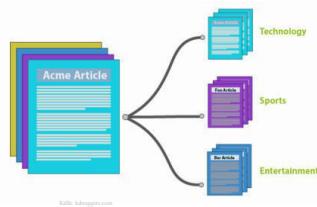
Naïve Bayes classifiers Decision trees

Lecture 2b

732A99/TDDE01

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



732A99/TDDE01

2

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

732A99/TDDE01

3

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!

732A99/TDDE01

4

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?
 - $P(\text{Space,ship}|\text{SciFi}) = P(\text{Space}|\text{SciFi}) * P(\text{Ship}|\text{SciFi})$?

732A99/TDDE01

5

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: $\hat{Y} = \arg \max_y p(Y = y|X)$

732A99/TDDE01

6

Naive Bayes classifiers - discrete inputs

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

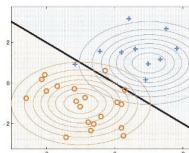
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

732A99/TDDE01

7

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 $p(x|y = C_i, \theta) = N(x|\mu_i, \Sigma_i)$
 - $\Sigma_i = \text{diag}(\sigma_{i1}^2, \dots, \sigma_{ip}^2)$
- Naive bayes is a special case of LDA (given A)**
 - MLE are means and variances (per class)

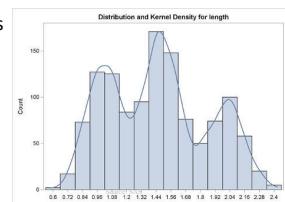


732A99/TDDE01

8

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



732A99/TDDE01

9

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)
```

732A99/TDDE01

10

Decision trees

Idea

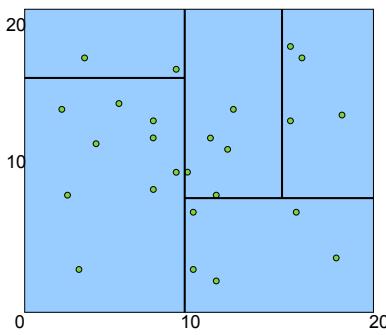
Split the domain of feature set into the set of hypercubes (rectangles, cubes) and define the target value to be constant within each hypercube

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

732A99/TDDE01

11

Classification tree toy example

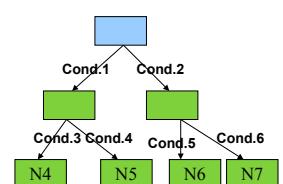


732A99/TDDE01

12

Definitions

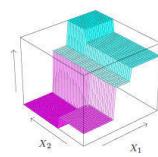
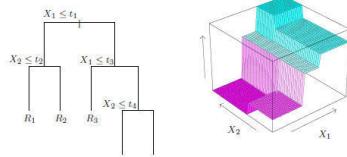
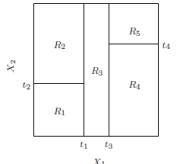
- Root node
- Nodes
- Leaves (terminal nodes)
- Parent node, child node
- Decision rules
- A value is assigned to the leaves



732A99/TDDE01

13

Regression tree toy example



732A99/TDDE01

14

A classification problem

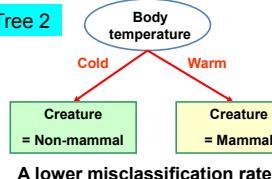
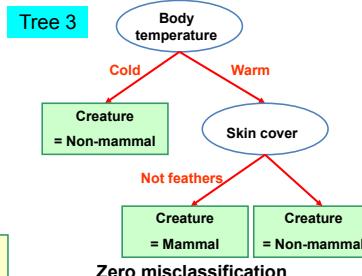
Create a classification tree that would describe the following patterns

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	yes	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	yes	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

732A99/TDDE01

15

Several solutions



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

732A99/TDDE01

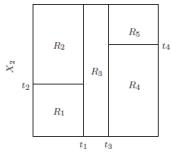
16

Decision trees

- A tree $T = \langle r_i, s_{r_i}, R_j, i = 1 \dots S, j = 1 \dots L \rangle$
 - $x_{r_i} \leq s_{r_i}$ 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



732A99/TDDE01

17

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

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

732A99/TDDE01

18

732A99/TDDE01

19

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 $Y_j | T \sim \text{Multinomial}(p_{j1}, \dots p_{jc})$

732A99/TDDE01

20

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
 $R_1(j, s) = \{X | X_j \leq s\}$ 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)

732A99/TDDE01

21

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?

732A99/TDDE01

22

Optimal trees

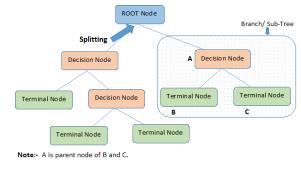
• Postpruning

Weakest link pruning:

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

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

 $|T| = \# \text{leaves}$
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!

732A99/TDDE01

23

Decision trees: comments

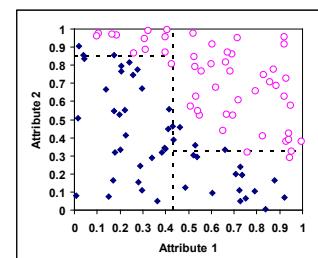
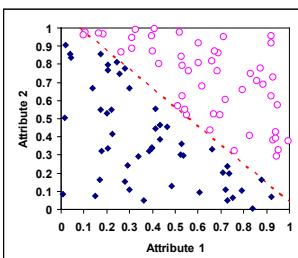
- Similar algorithms work for regression trees – replace $N \cdot Q(R)$ by $SSE(R)$
- Easy to interpret
- Easy to handle all types of features in one model
- **Automatic variable selection**
- Relatively robust to outliers
- Handle large datasets
- Trees have high variance: a small change in response → totally different tree
- Greedy algorithms → fit may be not so good
- Lack of smoothness

732A99/TDDE01

24

Decision trees: issues

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



732A99/TDDE01

25

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)
```

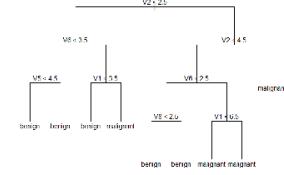
732A99/TDDE01

26

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

```
> fit
node * denotes terminal node
```

```
1) root 683 884 400 benign ( 0.65073 0.34992 )
  2) V2 < 2.5 418 108 900 benign ( 0.971292 0.028705 )
    4) V8 < 3.5 390 25 130 benign ( 1.000000 0.000000 )
      8) V6 < 3.0 380 13 100 benign ( 1.000000 0.000000 ) *
      9) V5 > 4.5 6 7 630 benign ( 0.666667 0.333333 ) *
    5) V8 > 3.5 23 31 490 benign ( 0.365212 0.434787 )
    10) V6 > 3.5 11 11 300 benign ( 0.365212 0.434787 ) *
     20) V1 < 3.5 11 11 300 benign ( 0.365212 0.434787 ) *
     11) V1 > 3.5 10 810 malignant ( 0.166667 0.833333 ) *
    3) V8 < 2.5 33 21 300 malignant ( 0.388889 0.611111 )
    6) V2 < 2.5 90 30 300 malignant ( 0.388889 0.611111 )
   12) V6 < 2.5 30 19 27.030 benign ( 0.833333 0.166667 )
   13) V8 < 2.5 19 15 15.000 benign ( 0.833333 0.166667 ) *
    24) V8 < 2.5 19 15 15.000 benign ( 0.833333 0.166667 ) *
    25) V8 > 2.5 19 15 15.000 malignant ( 0.545455 0.454545 ) *
   13) V6 > 2.5 60 54.070 malignant ( 0.166667 0.833333 )
   26) V1 < 6.5 28 33 160 malignant ( 0.365212 0.634787 ) *
    27) V1 > 6.5 90 90 90 malignant ( 0.365212 0.634787 ) *
   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
```

732A99/TDDE01

27

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
```

732A99/TDDE01

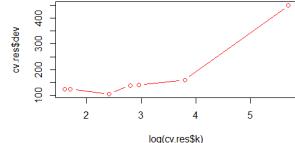
Decision trees in R

- Selecting optimal tree by penalizing

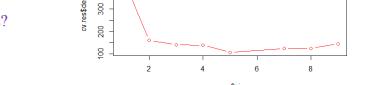
 - Cv.tree()

```
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$b), cv.res$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?

732A99/TDDE01

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
```

732A99/TDDE01

31

Generalized Linear Models: Uncertainty estimation

Lecture 2c

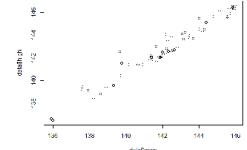
Moving beyond typical distributions

- We know how to model
 - Normally distributed targets \rightarrow linear regression
 - Bernoulli and Multinomial targets \rightarrow 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 \rightarrow multinomial does not work... (Poisson)

732A99/TDDE01

2

Exponential family

- More advanced error distributions are sometimes needed!
 - Many distributions belong to **exponential** family:
 - Normal, Exponential, Gamma, Beta, Chi-squared..
 - Bernoulli, Multinoulli, Poisson...
- $$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})e^{(\boldsymbol{\eta}^T u(\mathbf{x}))}$$
- Easy to find MLE and MAP
 - Non-exponential family distributions: uniform, Student t

Example: Bernoulli

732A99/TDDE01

3

Generalized linear models

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

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

- If f is a canonical link, then

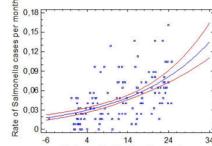
$$p(y|\mathbf{w}, s) = h(y, s)g(\mathbf{w}, \mathbf{x})e^{\frac{(\mathbf{w}^T \mathbf{x})y}{s}}$$

732A99/TDDE01

4

Generalized linear models

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



Example: Poisson regression

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

732A99/TDDE01

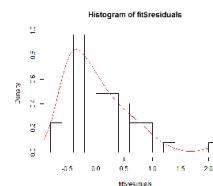
Generalized linear model: software

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

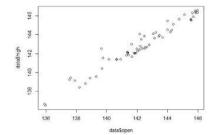
Example: Daily Stock prices NASDAQ

- Open
- High (within day)

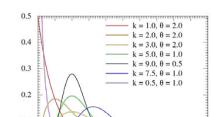
- Try to fit usual linear regression, study histogram of residuals



732A99/TDDE01



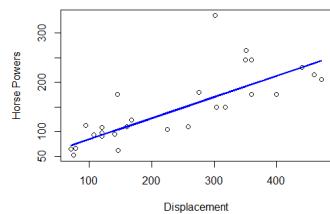
Gamma distribution: Wikipedia



6

Least absolute deviation regression

- Model $Y \sim \text{Laplace}(w^T X, b)$
 - 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**



732A99/TDDE01

7

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

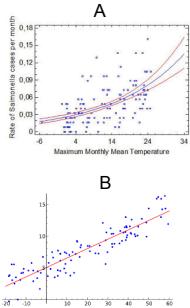
732A99/TDDE01

8

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 ..." 😊



732A99/TDDE01

9

Probabilistic models

Probabilistic model
 $Y \sim \text{Distribution}(f(x, w), \theta)$

- 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

732A99/TDDE01

10

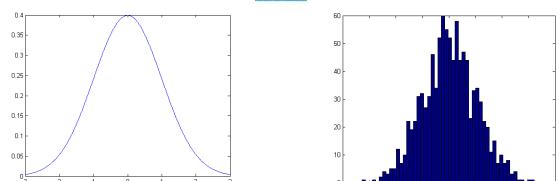
Uncertainty estimation

- Given estimator $\hat{f} = \hat{f}(x, D)$ (or $\hat{\alpha} = \delta(D)$), 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**

732A99/TDDE01

11

The bootstrap: general principle



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

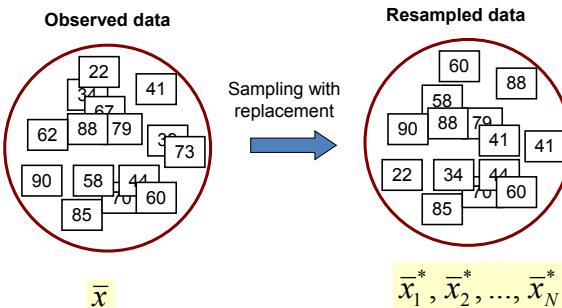
- Generate many different D_i from their distribution
- Use histogram of $\hat{f}(D_i, X)$ to determine confidence limits → unfortunately can not be done (distr of D is often unknown)

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

732A99/TDDE01

12

Nonparametric bootstrap



732A99/TDDE01

13

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 $D = (X_1, \dots, X_n)$
2. Generate $D_1 = (X_1^*, \dots, X_n^*)$ by sampling with replacement
3. Repeat step 2 B times
4. The distribution of w is given by $\hat{f}(D_1), \dots, \hat{f}(D_B)$

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

732A99/TDDE01

14

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

Parametric bootstrap is more precise if the distribution form is correct

732A99/TDDE01

15

Uncertainty estimation

1. Get D_1, \dots, D_B by bootstrap

2. Use $\hat{f}(D_1), \dots, \hat{f}(D_B)$ to estimate the uncertainty

- Bootstrap 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

732A99/TDDE01

16

Bootstrap confidence intervals

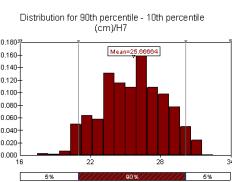
- To estimate $100(1-\alpha)$ confidence interval for w

Bootstrap percentile method

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

$$(w_{A_1}, w_{A_2})$$

Look at the plot...



732A99/TDDE01

17

Bootstrap: regression context

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

Nonparametric bootstrap:

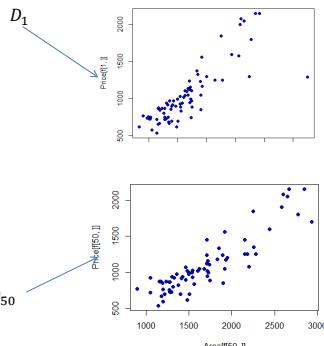
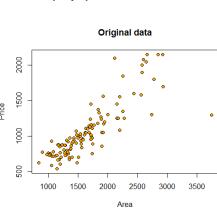
1. Using observation set D , sample pairs (X_i, Y_i) with replacement and get bootstrap sample D_1
2. Repeat step 1 B times → get D_1, \dots, D_B

732A99/TDDE01

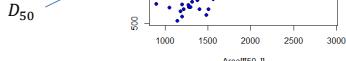
18

Uncertainty estimation

Example: Albuquerque dataset:
 $Y = \text{Price of House}$
 $X = \text{Area (sqft)}$



We sample data index, from $\{1 \dots N\}$



732A99/TDDE01

19

Bootstrap: regression context

Parametric bootstrap

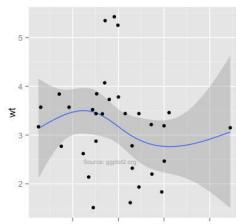
1. Fit a model to $D \rightarrow$ get $\hat{w}(D)$.
2. Set $X_i^* = X_i$, generate $Y_i^* \sim F(X_i, \hat{w})$.
3. $D_i = \{(X_i^*, Y_i^*), i = 1, \dots, n\}$
4. Repeat step 2 B times

732A99/TDDE01

20

Confidence intervals in regression

- Given $Y \sim \text{Distribution}(y|x, w)$, $EY|X = \mu|x = f(x, w)$
 - 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?
 - $\mu|x = f(x, w)$



732A99/TDDE01

21

Confidence intervals in regression

Estimation

1. Compute D_1, \dots, D_B using a bootstrap
2. Fit model to $D_1, \dots, D_B \rightarrow$ 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

732A99/TDDE01

22

Bootstrap: R

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

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

732A99/TDDE01

23

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
```

732A99/TDDE01

24

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

732A99/TDDE01

25

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")
```

732A99/TDDE01

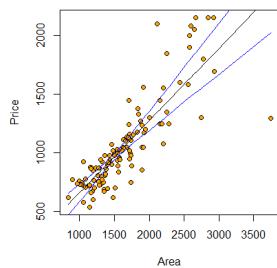
26

Uncertainty estimation: R

- Bootstrap confidence 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$Area,priceP,type="l") #plot fitted line
#plot confidence bands
points(data2$Area,e$point[2], type="l", col="blue")
points(data2$Area,e$point[1], type="l", col="blue")
```



732A99/TDDE01

27

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 $Distribution(X, \hat{w}(D_1)), \dots, Distribution(X, \hat{w}(D_B))$ and apply percentile method
4. Connect the intervals → get the band

732A99/TDDE01

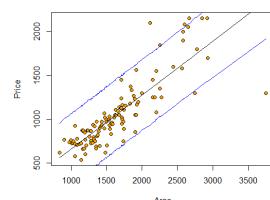
28

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?

732A99/TDDE01

29

Lecture 2d

Latent variable models

Overview

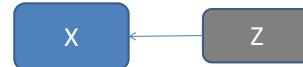
- Principal Component Analysis (PCA)
- Probabilistic PCA
- Independent component analysis (ICA)

732A99/TDDE01

2

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



732A99/TDDE01

3

Latent variables

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

3 | 3 | 3 | 3 | 3

732A99/TDDE01

4

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

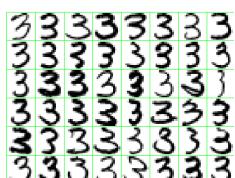
- Image recognition
- Information compression
- Subspace clustering
- ...

732A99/TDDE01

5

Principal Component Analysis (PCA)

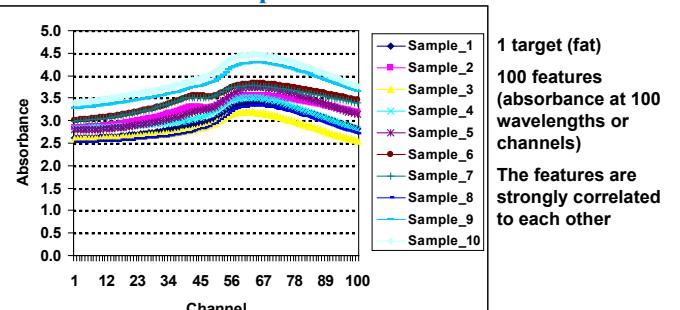
- Example 1: Handwritten digits
 - Can we get a more compact summary?



732A99/TDDE01

Absorbance records for ten samples of chopped meat

Parallel coordinate plot for “FAT”



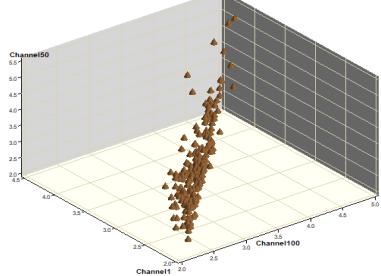
6

732A99/TDDE01

7

3-D plots of absorbance records for samples of meat - channels 1, 50 and 100

Scatterplot for three components



732A99/TDDE01

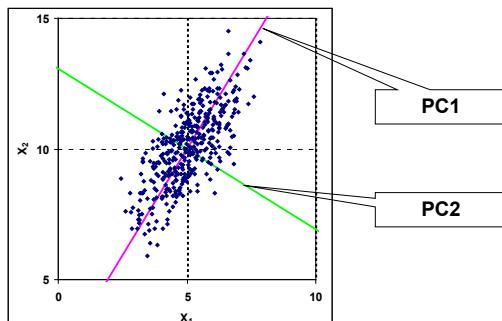
Principal components analysis

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
-

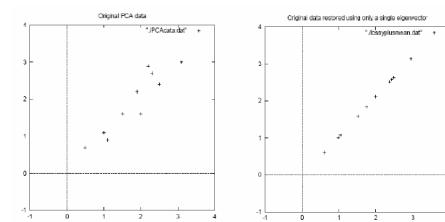
In the new coordinate system, coordinates corresponding to the last principal components are very small → can take away these columns

Principal Component Analysis - two inputs



732A99/TDDE01

PCA- after reducing dimensionality

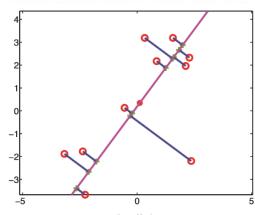


- Data became approximate (but less data to store)
- PC_1, \dots, PC_M are actually eigenvectors of sample covariance (first largest eigenvalue,...,Mth largest eigenvalue)

PCA: another view

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

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



732A99/TDDE01

PCA: computations

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

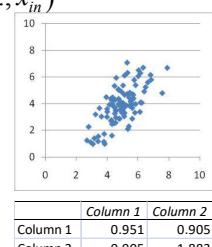
- Centred data

$$X = \|\mathbf{x}_1 - \bar{\mathbf{x}}_1 \mathbf{x}_2 - \bar{\mathbf{x}}_2 \dots \mathbf{x}_p - \bar{\mathbf{x}}_p\|,$$

- Covariance matrix

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

- Search for eigenvectors and eigenvalues of \mathbf{S}



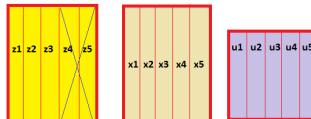
12

732A99/TDDE01

13

PCA: computations

4. Coordinates of any data point
 $x = (x_1 \dots x_p)$ 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

Store: $N \times M + p \times M$ instead $N \times p$

6. New data will have dimensions $N \times M$ instead of $N \times p$
 Getting approximate original data:

$$X' = ZU_M^T$$

732A99/TDDE01

14

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)

732A99/TDDE01

15

732A99/TDDE01

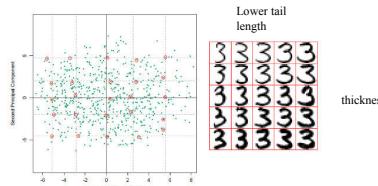
14

Principal Component Analysis

- Digits: two eigenvectors extracted

$$x = \boxed{3} + z1 \cdot \boxed{3} + z2 \cdot \boxed{3}$$

- Interpretation of eigenvectors



732A99/TDDE01

16

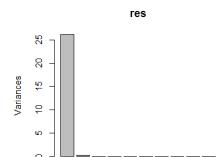
PCA in R

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

```
mydata=read.csv("tecator.csv")
data1$fat=c()
res=prcomp(data1)
lambda=res$dev^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-01
[7] 2.052212e-04 1.084213e-04 2.077326e-05 1.150359e-05

> 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!

732A99/TDDE01

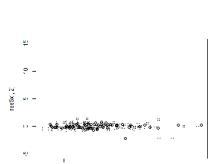
17

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 0.08184086 0.1227401 0.07125048 0.1156
```



Do we need second dimension?

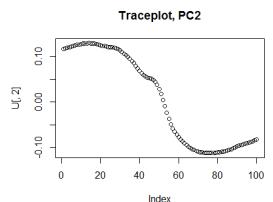
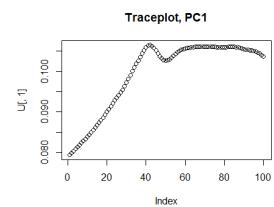
- Data in (PC1, PC2) – scores (Z)

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

732A99/TDDE01

- Trace plots

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



Which components contribute to PC1-2?

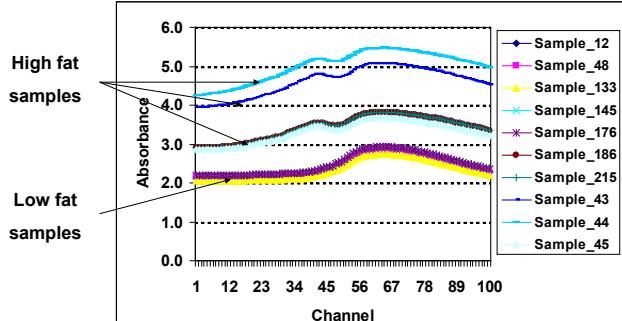
732A99/TDDE01

19

18

Absorbance records for ten samples of chopped meat

PCA2 captures the most of remaining variation



732A99/TDDE01

20

732A99/TDDE01

21

PCA for high-dimensional data

- Standard PCA for $p > N$

- At most N eigenvalues are nonzero
- Running time is $O(p^3)$

- High-dimensional PCA

1. Use $S' = \frac{1}{N} XX^T$ (instead of $S = \frac{1}{N} X^T X$)
2. Eigenvalues do not change
3. Eigenvectors of S are $X^T v_i$

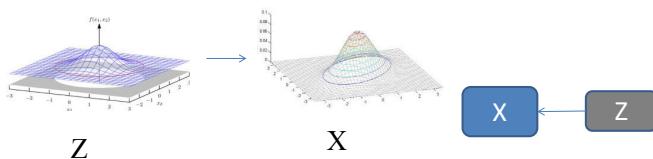
Probabilistic PCA

- z_i -latent variables, x_i - observed variables

$$z \sim N(0, I)$$

$$x|z \sim N(x|Wz + \mu, \sigma^2 I)$$
- Alternatively

$$z \sim N(0, I), x = \mu + Wz + \epsilon, \epsilon \sim N(0, \sigma^2 I)$$
- **Interpretation:** Observed data (X) is obtained by rotation, scaling and translation of standard normal distribution (Z) and adding some noise.



732A99/TDDE01

22

Probabilistic PCA

- **Aim:** extract Z from X

- Distribution of x :

$$x \sim N(\mu, C)$$

$$C = WW^T + \sigma^2 I$$

- Rotation invariance

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

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

- **Model will not be able find latent factors uniquely!** ⓘ

- It does not distinguish z from z'

732A99/TDDE01

23

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

732A99/TDDE01

24

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 $R = I, \sigma^2 = 0 \rightarrow$ get standard PCA components scaled by inverse root of eigenvalues

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

732A99/TDDE01

25

Advantages of probabilistic PCA

- More settings to specify → more flexible
- Can be faster when $M \ll p$
- 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

732A99/TDDE01

26

Probabilistic PCA in R

- Use **pcaMethods** from Bioconductor
 - Install
 - source("https://bioconductor.org/biocLite.R")
 - biocLite("pcaMethods")
- Ppca(data, nPcs,...)
- Results:** scores, loadings...

732A99/TDDE01

27

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 $p_i(z) = p(z)$
- Assuming latent features are **independent**

$$p(z) = \prod_{i=1}^M p_i(z_i)$$

732A99/TDDE01

28

ICA

- Setting $G_j(z) = -\log(p_j(z))$, $z_j = v_i^T x$ and assuming large sample

$$\min_V \sum_{j=1}^p E(G_j(z_j))$$

Subject to $\|v_i\| = 1$

Prewhitening

- Use PCA: $X' = XU$
- Computing z_i s for given V : $Z = X'V$

732A99/TDDE01

30

ICA

- Model

$$x = \mu + Wz + \epsilon, \quad \epsilon \sim N(0, \Sigma)$$
- **Estimation A: Maximum likelihood** ($V = W^{-1}$)
 - Assuming noise-free x

$$\max_V \sum_{i=1}^n \sum_{j=1}^p \log(p_j(v_j^T x_i))$$

Subject to $\|v_i\| = 1$

732A99/TDDE01

29

ICA

- **Estimation B: maximize negentropy**
 - ICA looks for model which is as much non-Gaussian as possible

$$\bullet \text{ Entropy } H(z) = -\int p(z) \log p(z) dz = E(-\log p(z))$$

$$\bullet \text{ Negentropy } J(z_i) = H(z'_i) - H(z_i)$$

– $z'_i \sim N(Ez_i, var(z_i))$

Negentropy maximization

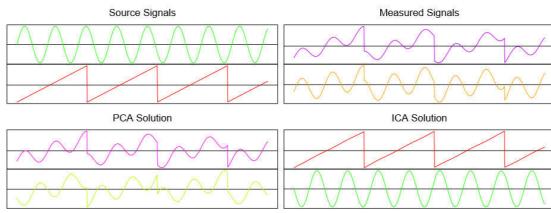
$$\max_V \sum_{j=1}^p J(z_j) = \min_V \sum_{j=1}^p H(z_j) = \min_V \sum_{j=1}^p E(-\log p(z_j))$$

732A99/TDDE01

31

ICA

- Example

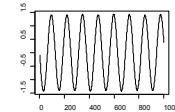
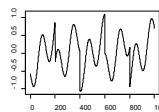
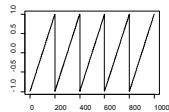
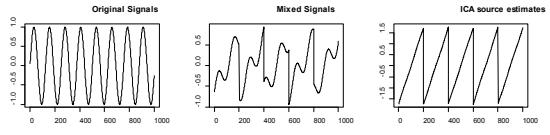


Source: Elm of stat learn by Tihon

732A99/TDDE01

Independent component analysis: R

```
S <- cbind(sin((1:1000)/20), rep(((1:200)-100)/100), 5)) #Generate data
A <- matrix(c(0.291, 0.6557, -0.5439, 0.5572), 2, 2)
X <- S %*% A
a <- fastICA(X, 2, alg.typ = "parallel", fun = "logcosh", alpha = 1,
method = "R", row.norm = FALSE, maxit = 200, tol = 0.0001, verbose = TRUE) #ICA
```



732A99/TDDE01

33

732A99/TDDE01 Machine Learning

Lecture 3a Block 1: Kernel Methods

Jose M. Peña
IDA, Linköping University, Sweden

1/19

Histogram Classification

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

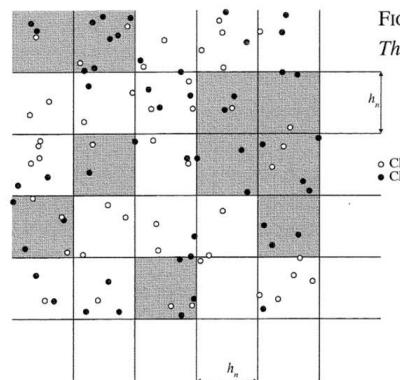


FIGURE 6.1. A cubic histogram rule:
The decision is 1 in the shaded area.

- In other words,

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

4/19

Literature

- Main source
 - Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006. Sections 2.5 and 6.1-6.2.
- Additional source
 - Devroye, L., Györfi, L. and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996. Sections 6.4 and 10.0.
 - Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning*. Springer, 2009. Chapter 6.

3/19

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.

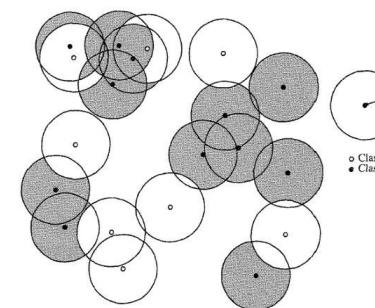


FIGURE 10.1. The moving window rule in \mathbb{R}^2 . The decision is 1 in the shaded area.

- In other words,

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

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

5/19

Kernel Classification

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

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

where $k : \mathbb{R}^D \rightarrow \mathbb{R}$ 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.

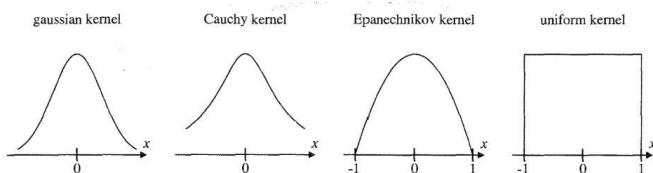


FIGURE 10.3. Various kernels on \mathcal{R} .

- 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: $k(u) = \mathbf{1}_{\{u \in S(0,1)\}}$

6/19

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 $y^*(\mathbf{x}) = \mathbb{E}_Y[y|\mathbf{x}]$.
- Since \mathbf{x} may not appear in the finite training set $\{(\mathbf{x}_n, t_n)\}$ available, then we average over the points in $C(\mathbf{x}, h)$ or $S(\mathbf{x}, h)$, or kernel-weighted average over all the points.
- In other words,

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

or

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

or

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

Kernel Classification

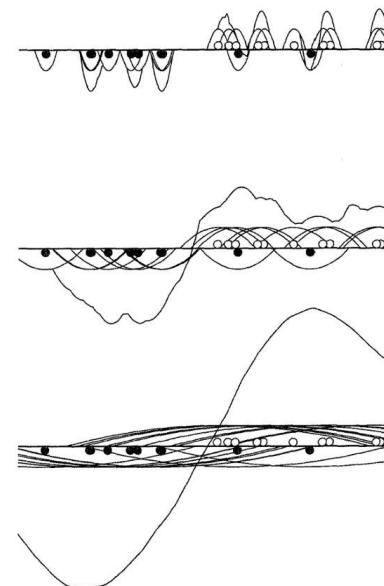


FIGURE 10.2. Kernel rule on the real line. The figure shows $\sum_{i=1}^n (2Y_i - 1) K((x - X_i)/h)$ for $n = 20$, $K(u) = (1 - u^2)\mathbf{1}_{\{|u| \leq 1\}}$ (the Epanechnikov kernel), and three smoothing factors h . One definitely undersmooths and one oversmooths. We took $p = 1/2$, and the class-conditional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

7/19

Histogram, Moving Window, and Kernel Density Estimation

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

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

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

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

and thus

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

- Then,

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

or

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

or

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

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

9/19

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 .

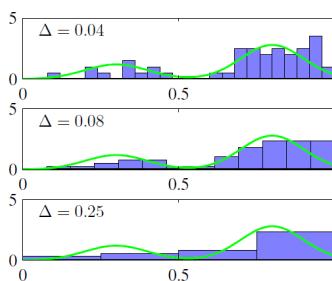
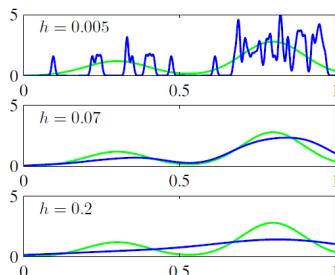


Figure 2.25 Illustration of the kernel density model (2.250) applied to the same data set used to demonstrate the histogram approach in Figure 2.24. We see that h acts as a smoothing parameter and that if it is set too small (top panel), the result is a very noisy density model, whereas if it is set 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).



10/19

Histogram, Moving Window, and Kernel Density Estimation

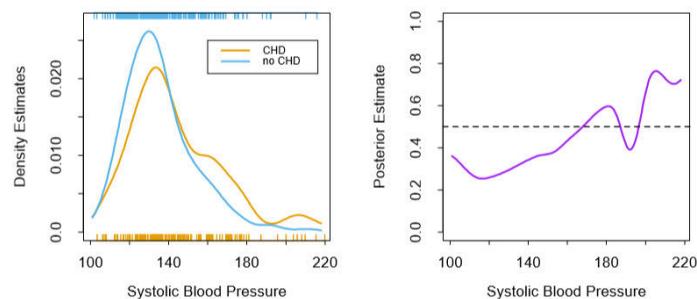


FIGURE 6.14. The left panel shows the two separate density estimates for systolic blood pressure in the CHD versus no-CHD groups, using a Gaussian kernel density estimate in each. The right panel shows the estimated posterior probabilities for CHD, using (6.25).

12/19

Histogram, Moving Window, and Kernel Density Estimation

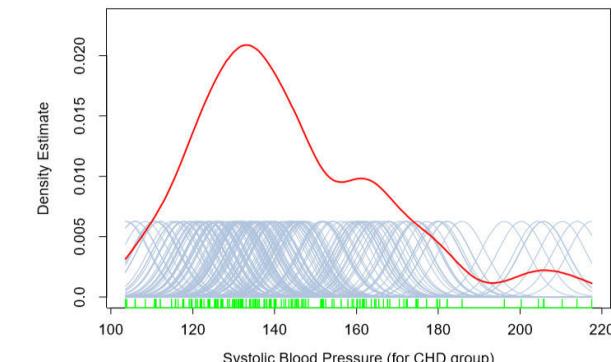


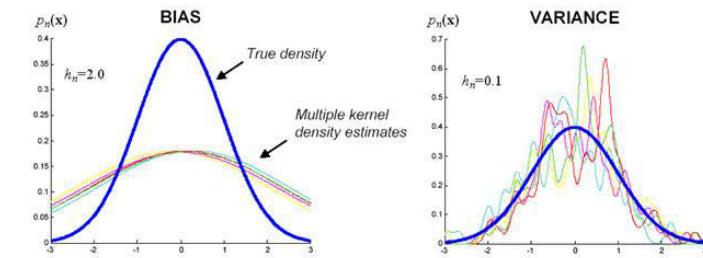
FIGURE 6.13. A kernel density estimate for systolic blood pressure (for the CHD group). The density estimate at each point is the average contribution from each of the kernels at that point. We have scaled the kernels down by a factor of 10 to make the graph readable.

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

11/19

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.

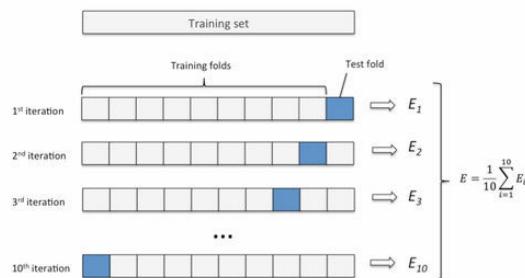


- ▶ Small width implies considering few points. So, the variance will be large (similar to the variance of a single point). The bias will be small since the points considered are close to \mathbf{x} .
- ▶ Large width implies considering many points. So, the variance will be small and the bias will be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

13/19

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.

14/19

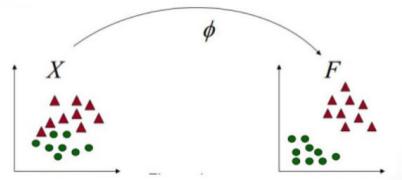
Kernel Trick

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

- If the matrix

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

is symmetric and positive semi-definite for all choices of $\{\mathbf{x}_n\}$, then $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$ 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,

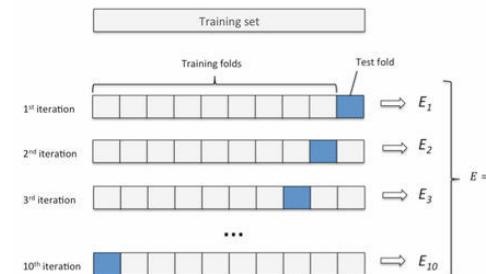
$$\phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2c}x_1, \sqrt{2c}x_2, c)$$

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

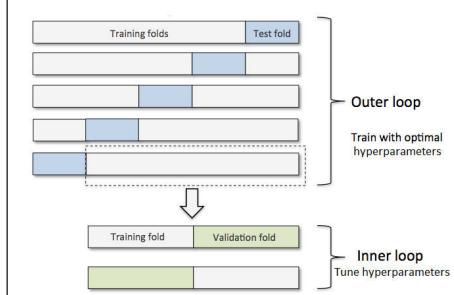
Kernel Selection

- 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 ?

15/19

Kernel Trick

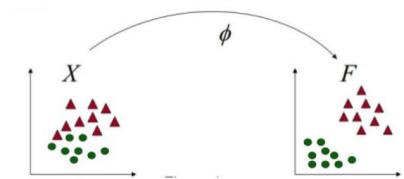
- Consider again moving window classification, regression, and density estimation.
- Note that $\mathbf{x}_n \in S(\mathbf{x}, h)$ if and only if $\|\mathbf{x} - \mathbf{x}_n\| \leq h$.
- Note that

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

- Then,

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

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



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

17/19

Kernel Trick

- Two alternatives for building $k(\mathbf{x}, \mathbf{x}')$:
 - Choose a convenient $\phi(\mathbf{x})$ and let $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$.
 - Build it from existing kernel functions as follows.

Techniques for Constructing New Kernels.

Given valid kernels $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$, the following new kernels will also be valid:

$$k(\mathbf{x}, \mathbf{x}') = ck_1(\mathbf{x}, \mathbf{x}') \quad (6.13)$$

$$k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') \quad (6.14)$$

$$k(\mathbf{x}, \mathbf{x}') = q(k_1(\mathbf{x}, \mathbf{x}')) \quad (6.15)$$

$$k(\mathbf{x}, \mathbf{x}') = \exp(k_1(\mathbf{x}, \mathbf{x}')) \quad (6.16)$$

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') \quad (6.17)$$

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') \quad (6.18)$$

$$k(\mathbf{x}, \mathbf{x}') = k_3(\phi(\mathbf{x}), \phi(\mathbf{x}')) \quad (6.19)$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{A} \mathbf{x}' \quad (6.20)$$

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b) \quad (6.21)$$

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a)k_b(\mathbf{x}_b, \mathbf{x}'_b) \quad (6.22)$$

where $c > 0$ is a constant, $f(\cdot)$ is any function, $q(\cdot)$ is a polynomial with nonnegative coefficients, $\phi(\mathbf{x})$ is a function from \mathbf{x} to \mathbb{R}^M , $k_3(\cdot, \cdot)$ is a valid kernel in \mathbb{R}^M , \mathbf{A} is a symmetric positive semidefinite matrix, \mathbf{x}_a and \mathbf{x}_b are variables (not necessarily disjoint) with $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$, and k_a and k_b are valid kernel functions over their respective spaces.

Summary

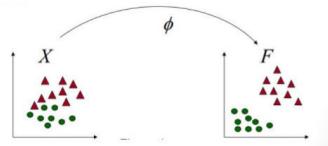
- Kernel methods: Smoothing models.
- Model selection: Nested cross-validation.
- Kernel trick: It allows to work in the feature space without constructing it.

Support Vector Machines for Classification

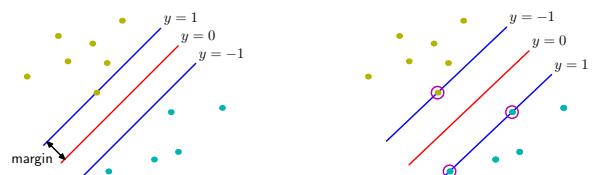
- Consider binary classification with input space \mathbb{R}^D .
- Consider a training set $\{(\mathbf{x}_n, t_n)\}$ where $t_n \in \{-1, +1\}$.
- Consider using the linear model

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

- so that a new point \mathbf{x} is classified according to the sign of $y(\mathbf{x})$.
- Assume that the training set is linearly separable in the feature space (but not necessarily in the input space), i.e. $t_n y(\mathbf{x}_n) > 0$ for all n .

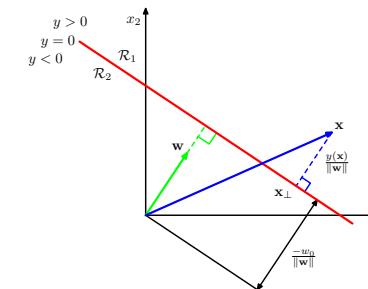


- Aim for the separating hyperplane that maximizes the margin (i.e. the smallest perpendicular distance from any point to the hyperplane) so as to minimize the generalization error.



18/19

Support Vector Machines for Classification



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

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

- Then, the maximum margin separating hyperplane is given by

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

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

4/18

19/19

Support Vector Machines for Classification

- Then, the maximum margin separating hyperplane is given by

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

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

- To minimize the previous expression, we minimize

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

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 \mathbf{w} and b , and maximizing with respect to a_n .
- Setting its derivatives with respect to \mathbf{w} and b to zero gives

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

$$0 = \sum_n a_n t_n$$

6/18

Support Vector Machines for Classification

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

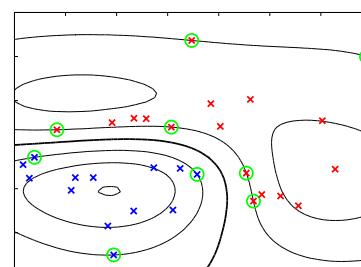
$$a_n(t_n y(\mathbf{x}_n) - 1) = 0$$

- Then, $a_n > 0$ if and only if $t_n y(\mathbf{x}_n) = 1$. The points with $a_n > 0$ are called support vectors and they lie on the margin boundaries.

- A new point \mathbf{x} is classified according to the sign of

$$\begin{aligned} y(\mathbf{x}) &= \mathbf{w}^T \phi(\mathbf{x}) + b = \sum_n a_n t_n \phi(\mathbf{x}_n)^T \phi(\mathbf{x}) + b = \sum_n a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b \\ &= \sum_{m \in S} a_m t_m k(\mathbf{x}, \mathbf{x}_m) + b \end{aligned}$$

where S are the indexes of the support vectors. Sparse solution!



8/18

Support Vector Machines for Classification

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

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

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.

7/18

Support Vector Machines for Classification

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

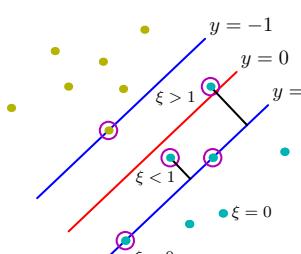
$$1 = t_n y(\mathbf{x}_n) = t_n \left(\sum_{m \in S} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b \right)$$

and multiplying both sides by t_n , we have that

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

- 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

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



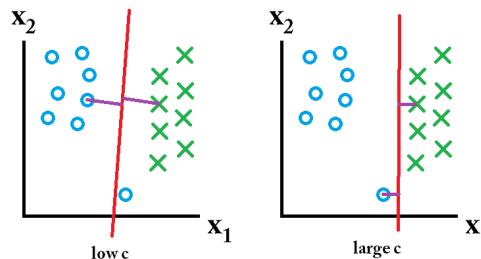
9/18

Support Vector Machines for Classification

- The optimal separating hyperplane is given by

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

subject to $t_n y(\mathbf{x}_n) \geq 1 - \xi_n$ 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

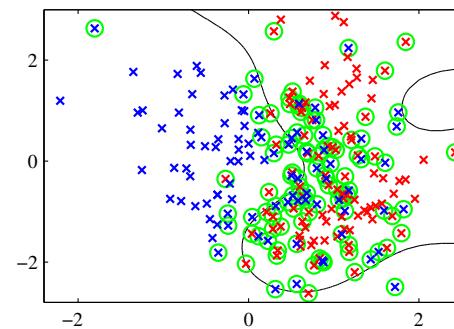
$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_n \xi_n - \sum_n a_n (t_n (\mathbf{w}^T \phi(\mathbf{x}_n) + b) - 1 + \xi_n) - \sum_n \mu_n \xi_n$$

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

10/18

Support Vector Machines for Classification

- Since the optimal \mathbf{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 \mathbf{x}_n with $0 < a_n < C$.



- Not covered topics:

- Classifying into more than two classes.
- Returning class posterior probabilities.

Support Vector Machines for Classification

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

$$\begin{aligned} \mathbf{w} &= \sum_n a_n t_n \phi(\mathbf{x}_n) \\ 0 &= \sum_n a_n t_n \\ a_n &= C - \mu_n \end{aligned}$$

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

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

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 :

$$\begin{aligned} a_n(t_n y(\mathbf{x}_n) - 1 + \xi_n) &= 0 \\ \mu_n \xi_n &= 0 \end{aligned}$$

- Then, $a_n > 0$ if and only if $t_n y(\mathbf{x}_n) = 1 - \xi_n$ 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.

11/18

Support Vector Machines for Regression

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

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

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

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

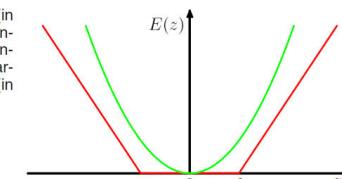
consider minimizing the ϵ -insensitive regularized error function

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

where $C > 0$ controls regularization and

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

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).



13/18

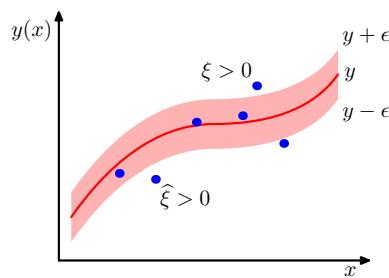
Support Vector Machines for Regression

- 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

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

and

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



14/18

Support Vector Machines for Regression

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

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

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 :

$$\begin{aligned} a_n(y(\mathbf{x}_n) + \epsilon + \xi_n - t_n) &= 0 \\ \widehat{a}_n(t_n - y(\mathbf{x}_n) + \epsilon + \widehat{\xi}_n) &= 0 \\ \mu_n \xi_n &= 0 \\ \widehat{\mu}_n \widehat{\xi}_n &= 0 \end{aligned}$$

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

16/18

Support Vector Machines for Regression

- The optimal regression curve is given by

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

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

- To minimize the previous expression, we minimize

$$\begin{aligned} & C \sum_n (\xi_n + \widehat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_n (\mu_n \xi_n + \widehat{\mu}_n \widehat{\xi}_n) \\ & - \sum_n a_n (y(\mathbf{x}_n) + \epsilon + \xi_n - t_n) - \sum_n \widehat{a}_n (t_n - y(\mathbf{x}_n) + \epsilon + \widehat{\xi}_n) \end{aligned}$$

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 \mathbf{w} , b , ξ_n and $\widehat{\xi}_n$ to zero gives

$$\mathbf{w} = \sum_n (a_n - \widehat{a}_n) \phi(\mathbf{x}_n)$$

$$0 = \sum_n (a_n - \widehat{a}_n)$$

$$C = a_n + \mu_n$$

$$C = \widehat{a}_n + \widehat{\mu}_n$$

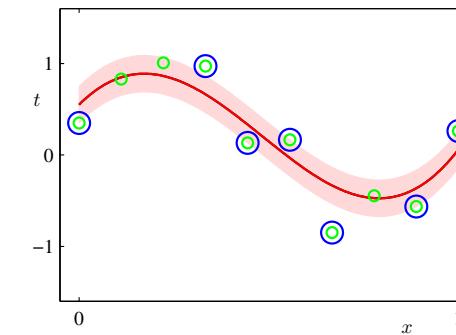
15/18

Support Vector Machines for Regression

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

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

where S are the indexes of the support vectors. Sparse solution!



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

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

17/18

- Kernel trick: It allows to work in the feature space without constructing it.
- Quadratic objective function: It allows to obtain the global optimum for a given kernel and C/ϵ (which are obtained by cross-validation).
- Sparse model: Only the support vectors are needed for classification/regression (compare with kernel models).

18/18

Neural Networks

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

$$y(\mathbf{x}) = \text{sgn} \left(\sum_{m \in S} a_m t_m k(\mathbf{x}, \mathbf{x}_m) + b \right)$$

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

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

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

4/16

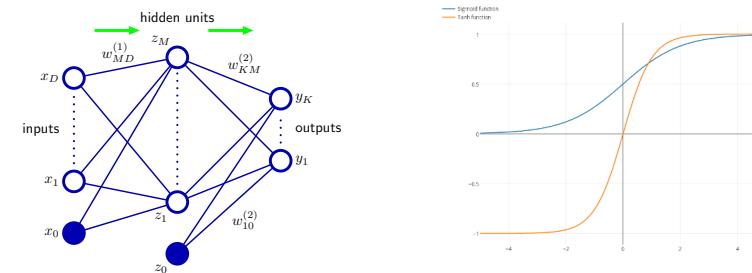
732A99/TDDE01 Machine Learning

Lecture 3c Block 1: Neural Networks

Jose M. Peña
IDA, Linköping University, Sweden

1/16

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: $y_k(\mathbf{x}) = a_k$
- Output activation function for classification: $y_k(\mathbf{x}) = \sigma(a_k)$
- Sigmoid function: $\sigma(a) = \frac{1}{1+\exp(-a)}$
- Two-layer NN:

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

- 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.

5/16

Neural Networks

- For a large variety of activation functions, the two-layer NN can uniformly approximate any continuous function to arbitrary accuracy provided enough hidden units. Easy to fit the parameters ? Overfitting ?!

Figure 5.3 Illustration of the capability of a multilayer perceptron to approximate four different functions comprising (a) $f(x) = x^2$, (b) $f(x) = \sin(x)$, (c) $f(x) = |x|$, and (d) $f(x) = H(x)$ where $H(x)$ is the Heaviside step function. In each case, $N = 50$ data points, shown as blue dots, have been sampled uniformly in x over the interval $(-1, 1)$ and the corresponding values of $f(x)$ evaluated. These data points are then used to train a two-layer network having 3 hidden units with 'tanh' activation functions and linear output units. The resulting network functions are shown by the red curves, and the outputs of the three hidden units are shown by the three dashed curves.

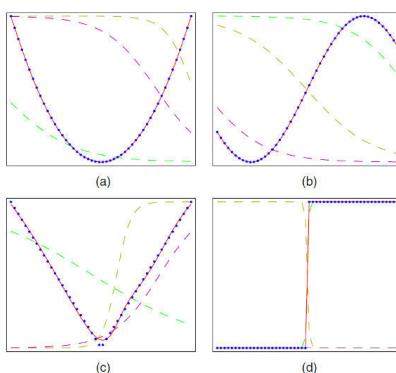
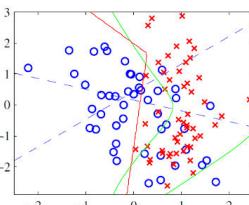


Figure 5.4 Example of the solution of a simple two-class classification problem involving synthetic data using a neural network having two inputs, two hidden units with 'tanh' activation functions, and a single output having a logistic sigmoid activation function. The dashed blue lines show the $z = 0.5$ contours for each of the hidden units, and the red line shows the $y = 0.5$ decision surface for the network. For comparison, the green line denotes the optimal decision boundary computed from the distributions used to generate the data.



6/16

Backpropagation Algorithm

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

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

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

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

- Then, the likelihood function is

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

and thus

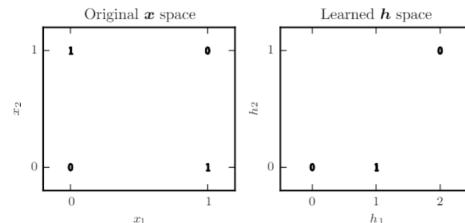
$$-\ln p(\{\mathbf{t}_n\} | \{\mathbf{x}_n\}, \mathbf{w}, \sigma) = \sum_n \sum_k \frac{1}{2\sigma^2} (t_{nk} - y_k(\mathbf{x}_n))^2 + \frac{N}{2} \ln \sigma^2 + \frac{N}{2} \ln 2\pi$$

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 \mathbf{w} , plug them into the log likelihood function, and maximize it with respect to σ .

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.



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

7/16

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

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

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

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

where n is chosen randomly or sequentially.

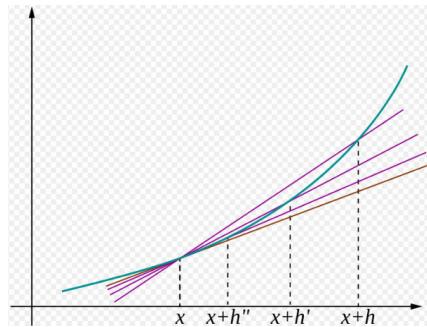
- Sequential gradient descent is less affected by the multimodality problem, as a local minimum of the whole data will not be generally a local minimum of each individual point.

8/16

9/16

Backpropagation Algorithm

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



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

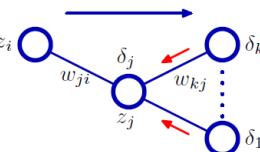
10/16

Backpropagation Algorithm

Backpropagation algorithm:

- Forward propagate to compute activations, and hidden and output units.
- Compute δ_k for the output units.
- Backpropagate the δ 's, i.e. evaluate δ_j for the hidden units recursively.
- 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:

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

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

12/16

Backpropagation Algorithm

- Since E_n depends on w_{ji} only via a_j , and $a_j = \sum_i w_{ji} x_i$, then

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{\partial E_n}{\partial a_j} \frac{\partial a_j}{\partial w_{ji}} = \frac{\partial E_n}{\partial a_j} x_i = \delta_j x_i$$

- Since E_n depends on a_j only via a_k , then

$$\delta_j = \frac{\partial E_n}{\partial a_j} = \sum_k \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial a_j} = \sum_k \delta_k \frac{\partial a_k}{\partial a_j}$$

- Since $a_k = \sum_j w_{kj} z_j$ and $z_j = h(a_j)$, then

$$\frac{\partial a_k}{\partial a_j} = h'(a_j) w_{kj}$$

- Putting all together, we have that

$$\delta_j = h'(a_j) \sum_k \delta_k w_{kj}$$

- Since $y_k = a_k$ for regression and $a_k = \sum_j w_{kj} z_j$, then

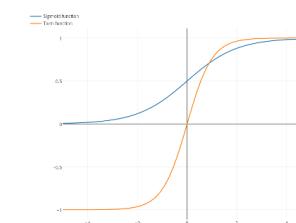
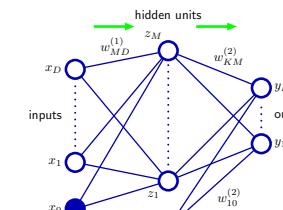
$$\frac{\partial E_n}{\partial w_{kj}} = \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial w_{kj}} = \delta_k z_j \text{ and } \delta_k = \frac{\partial E_n}{\partial a_k} = y_k - t_k$$

Backpropagation algorithm:

- Forward propagate to compute activations, and hidden and output units.
- Compute δ_k for the output units.
- Backpropagate the δ 's, i.e. evaluate δ_j for the hidden units recursively.
- Compute the required derivatives.

11/16

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:

- Forward propagation, i.e. compute

$$a_j = \sum_i w_{ji} x_i \text{ and } z_j = h(a_j) \text{ and } y_k = \sum_j w_{kj} z_j$$

- Compute

$$\delta_k = y_k - t_k$$

- Backpropagate, i.e. compute

$$\delta_j = (1 - z_j^2) \sum_k w_{kj} \delta_k$$

- Compute

$$\frac{\partial E_n}{\partial w_{kj}} = \delta_k z_j \text{ and } \frac{\partial E_n}{\partial w_{ji}} = \delta_j x_i$$

13/16

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.

14/16

Summary

- ▶ NNs: Nonlinear mapping from input to output.
- ▶ Extremely expressive.
- ▶ Training: Backpropagation algorithm, and regularization.

16/16

Regularization

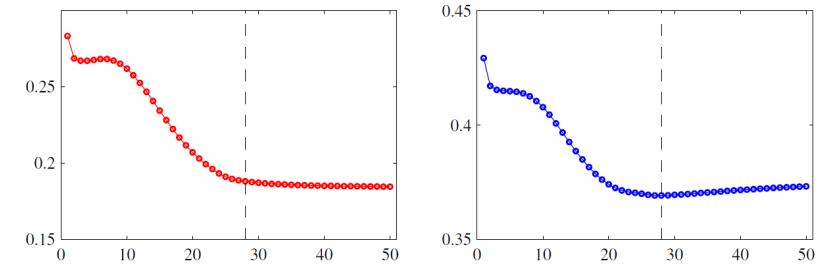


Figure 5.12 An illustration of the behaviour of training set error (left) and validation set error (right) during a typical training session, as a function of the iteration step, for the sinusoidal data set. The goal of achieving the best generalization performance suggests that training should be stopped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

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

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.

15/16

Limitations of Neural Networks

Theorem (Universal approximation theorem)

For every continuous function $f : [a, b]^D \rightarrow \mathbb{R}$ 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 $\mathcal{C}^{(k)}$ contain all classifiers defined by NNs of one hidden layer with k hidden units and the sigmoid activation function. Then, for any distribution $p(\mathbf{x}, t)$,

$$\lim_{k \rightarrow \infty} \inf_{y \in \mathcal{C}^{(k)}} \mathbb{E}_x[y(\mathbf{x})] - \mathbb{E}_x[p(t|\mathbf{x})] = 0$$

- ▶ 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.

4/16

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.

5/16

Deep Neural Networks

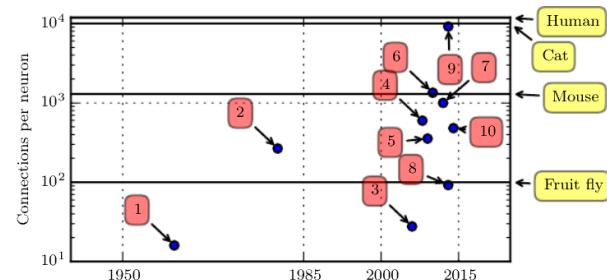


Figure 1.10: Initially, the number of connections between neurons in artificial neural networks was limited by hardware capabilities. Today, the number of connections between neurons is mostly a design consideration. Some artificial neural networks have nearly as many connections per neuron as a cat, and it is quite common for other neural networks to have as many connections per neuron as smaller mammals like mice. Even the human brain does not have an exorbitant amount of connections per neuron. Biological neural network sizes from [Wikipedia \(2015\)](#).

1. Adaptive linear element (Widrow and Hoff, 1960)
2. Neocognitron (Fukushima, 1980)
3. GPU-accelerated convolutional network (Chellapilla et al., 2006)
4. Deep Boltzmann machine (Salakhutdinov and Hinton, 2009a)
5. Unsupervised convolutional network (Jarrett et al., 2009)
6. GPU-accelerated multilayer perceptron (Ciresan et al., 2010)
7. Distributed autoencoder (Le et al., 2012)
8. Multi-GPU convolutional network (Krizhevsky et al., 2012)
9. COTS HPC unsupervised convolutional network (Coates et al., 2013)
10. GoogLeNet (Szegedy et al., 2014a)

7/16

Deep Neural Networks

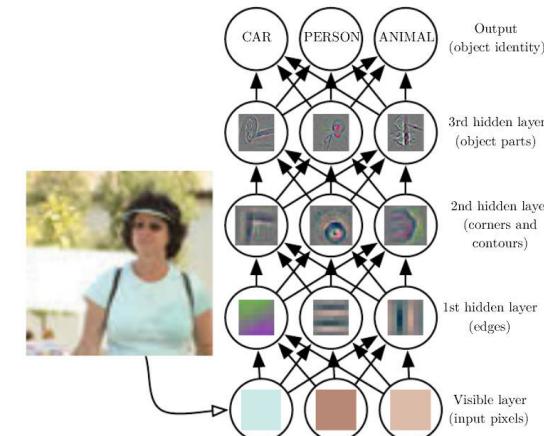


Figure 1.11: Since the introduction of hidden units, artificial neural networks have doubled in size roughly every 2.4 years. Biological neural network sizes from [Wikipedia \(2015\)](#).

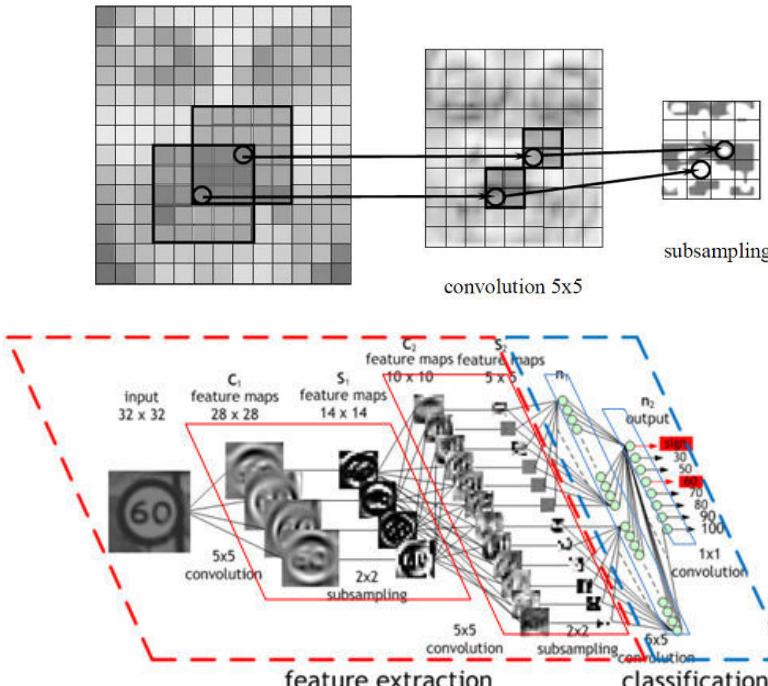
1. Perceptron (Rosenblatt, 1958, 1962)
 2. Adaptive linear element (Widrow and Hoff, 1960)
 3. Neocognitron (Fukushima, 1980)
 4. Early back-propagation network (Rumelhart et al., 1986b)
 5. Recurrent neural network for speech recognition (Robinson and Fallside, 1991)
 6. Multilayer perceptron for speech recognition (Bengio et al., 1991)
 7. Mean field sigmoid belief network (Saul et al., 1996)
 8. LeNet-5 (LeCun et al., 1998b)
 9. Echo state network (Jaeger and Haas, 2004)
 10. Deep belief network (Hinton et al., 2006)
 11. GPU-accelerated convolutional network (Chellapilla et al., 2006)
 12. Deep Boltzmann machine (Salakhutdinov and Hinton, 2009a)
 13. GPU-accelerated deep belief network (Raina et al., 2009)
 14. Unsupervised convolutional network (Jarrett et al., 2009)
 15. GPU-accelerated multilayer perceptron (Ciresan et al., 2010)
 16. OMP-1 network (Coates and Ng, 2011)
 17. Distributed autoencoder (Le et al., 2012)
 18. Multi-GPU convolutional network (Krizhevsky et al., 2012)
 19. COTS HPC unsupervised convolutional network (Coates et al., 2013)
 20. GoogLeNet (Szegedy et al., 2014a)
- 22 layers DNN, but
12 times fewer weights
than DNN 19

8/16

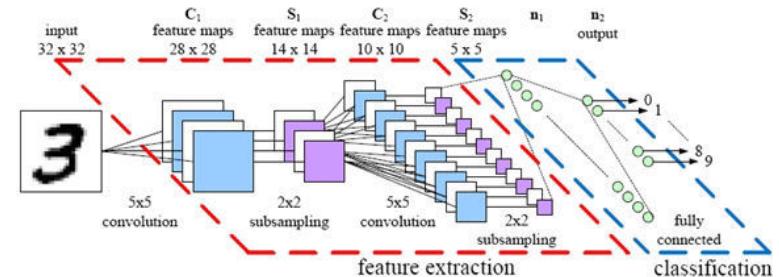
- ▶ Training DNNs is difficult:
 - ▶ Typically, poorer generalization than (shallow) NNs.
 - ▶ The gradient may vanish/explode as we move away from the output layer, due to multiplying small/big quantities. E.g. the gradient of σ and \tanh is in $[0, 1]$. So, they may only suffer the gradient vanishing problem. Other activation 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".

9/16

Convolutional Networks



11/16



- ▶ 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.

10/16

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.

12/16

Rectifier Activation Function

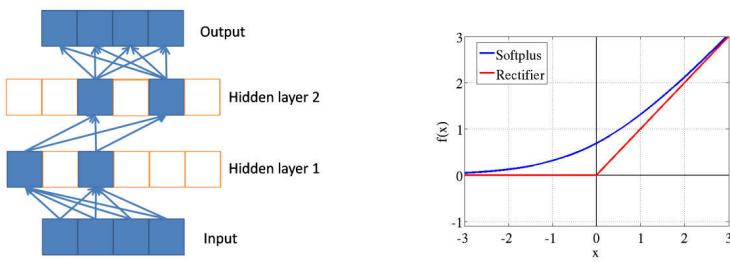


Figure 2: *Left:* Sparse propagation of activations and gradients in a network of rectifier units. The input selects a subset of active neurons and computation is linear in this subset. *Right:* Rectifier and softplus activation functions. The second one is a smooth version of the first.

- ▶ $\text{rectifier}(x) = \max\{0, x\}$, 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.

13/16

Layer-Wise Pre-Training

- ▶ The pre-training aims to find a good starting point for the subsequent run of the backpropagation algorithm.
- ▶ Supervised version:
 1. Train each layer of the DNN as if it was the hidden layer in a depth-two NN. As input, use the output of the last of the previously trained layers. As output, use the original classification or regression function.
 2. Run the backpropagation algorithm to fine-tune the weights.
- ▶ Unsupervised version: Similar to the supervised one but the hidden layers (except the last one) are trained to learn an encoding of the output of the previous layer, instead of the original classification or regression function.

15/16

Rectifier Activation Function

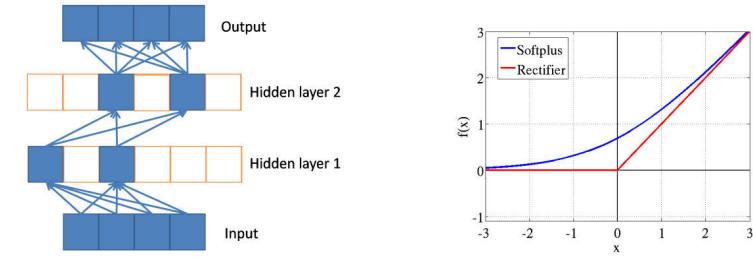


Figure 2: *Left:* Sparse propagation of activations and gradients in a network of rectifier units. The input selects a subset of active neurons and computation is linear in this subset. *Right:* Rectifier and softplus activation functions. The second one is a smooth version of the first.

- ▶ 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.

14/16

Summary

- ▶ Direct application of the backpropagation algorithm to DNNs produces poor results.
- ▶ Convolutional networks: It makes the backpropagation algorithm more efficient by using local features and weight sharing. This also achieves invariance, which is particularly important for image processing.
- ▶ Rectifier activation function: Free of gradient vanishing problem and it simplifies the backpropagation algorithm.
- ▶ Layer-wise pre-training: Heuristic weight initialization to alleviate the local optima problem.

16/16