Welcome ...

...to the course in Statistical Learning, 2011.

Lectures: Niels Richard Hansen

- Co-taught with a Statistical Learning course at the University of Copenhagen.
- Evaluation:
  A final, individual project handed with hand-in deadline via email Monday, October 31.
- Theoretical and practical exercises: During the course I plan to give an number of small theoretical exercises and practical (mostly `R`-) exercises that you will solve/work on in class or for the subsequent lecture. Solutions will be provided. Some of the exercises are taken from the book or are slightly modified versions of exercises from the book.
- Teaching material: *The Elements of Statistical Learning. Data Mining, Inference, and Prediction 2nd ed.* together with hand-outs from the lectures.

Statistical Learning

What is Statistical Learning?

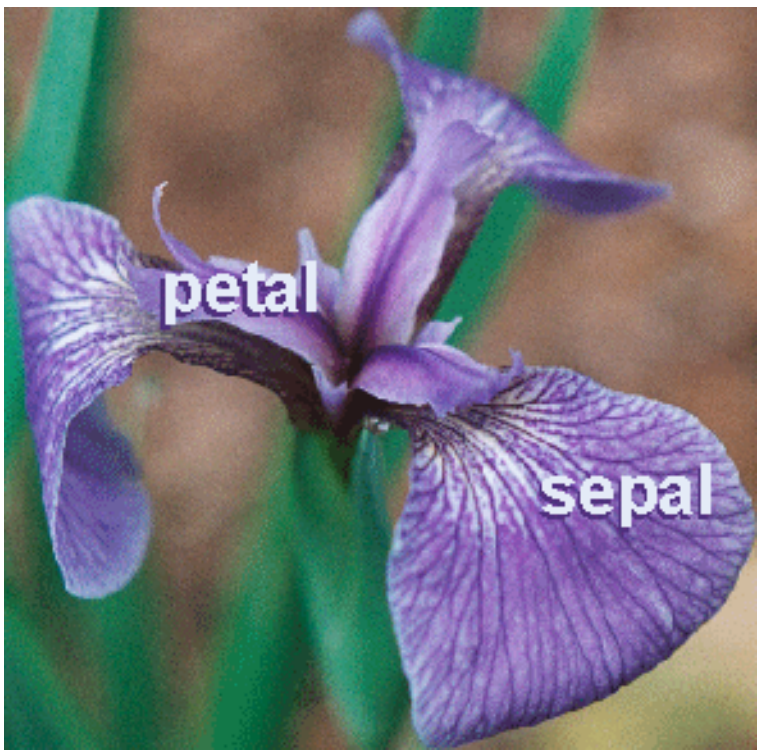
A merger of classical disciplines in statistics with methodology from areas known as machine learning, pattern recognition and artificial neural networks.

Major purpose: Prediction – as opposed to .... truth!?

Major point of view: Function approximation, solution of a mathematically formulated estimation problem – as opposed to algorithms.

Iris data

A classical dataset collected by the botanist Edgar Anderson, 1935, *The irises of the Gaspe Peninsula* and studied by statistician R. A. Fisher, 1936, *The use of multiple measurements in taxonomic problems*. Available as the `iris` dataset in the `datasets` package in `R`.



Sepal		Petal		
Length	Width	Length	Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
⋮	⋮	⋮	⋮	⋮
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
⋮	⋮	⋮	⋮	⋮
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
⋮	⋮	⋮	⋮	⋮

Figure 1.1 – Prostate Cancer

A classical scenario from statistics. How does the response variable `lpsa` relate to a number of other measured or observed quantities – some continuous and some categorical?

Typical approach is regression – the scatter plot might reveal marginal correlations.

Figure 1.2 – Hand Written Digits

A classical problem from pattern recognition. How do we classify an image of a handwritten number as 0 - 9?

This is the mail sorting problem based on zip codes.

It's not so easy – is a nine or a five?

Figure 1.3 – Microarray Measurements

A problem of current importance. How does the many genes of our cells behave?

We can measure the activity of thousands of genes simultaneously – the gene expression levels – and want to know about the relation of gene expression patterns to "status of the cell" (healthy, sick, cancer, what type of cancer ...)

Classification

The objective in a classification problem is to be able to classify an object into a finite number of distinct groups based on observed quantities.

With hand written digits we have 10 groups and an 16x16 pixel gray tone image (a vector in $\mathbb{R}^{256}$).

With microarrays a typical scenario is that we have 2 groups (cancer type A and cancer type B) and a 10-30 thousand dimensional vector of gene expressions.

Setup – and One Simple Idea

We have observations $(x_1, y_1), \ldots, (x_N, y_N)$ with $x_i \in \mathbb{R}^p$ and $y_i \in \{0, 1\}$. We assume that the data arose as independent and identically distributed samples of a pair (X, Y) of random variables.

Assume $X = x_0 \in \mathbb{R}^p$ what is Y? Let

$$N_k(x_0) = \{i \mid x_i \text{ is one of the } k\text{'th nearest observations}\}.$$

Define

$$\hat{p}(x_0) = \frac{1}{k} \sum_{i \in N_k(x_0)} y_i \in [0, 1]$$

and classify using majority rules

$$\hat{y} = \hat{f}(x) = \begin{cases} 1 & \text{if } \hat{p}(x_0) \geq 1/2 \\ 0 & \text{if } \hat{p}(x_0) < 1/2 \end{cases}$$

Figure 2.2 – 15-Nearest Neighbor Classifier

A wiggly separation barrier between x_0's classified as zero's and one's is characteristic of nearest neighbors. With $k = 15$ we get a partition of the space into just two connected "classification components".

Figure 2.3 – 1-Nearest Neighbor Classifier

With $k = 1$ every observed point has its own "neighborhood of classification". The result is a large(r) number of connected classification components.

Linear Classifiers

A classifier is called linear if there is an affine function

$$x \mapsto x^T\beta + \beta_0$$

with the classifier at x_0

$$f(x) = \begin{cases} 1 & \text{if } x^T\beta + \beta_0 \geq 0 \\ 0 & \text{if } x^T\beta + \beta_0 < 0 \end{cases}$$

There are several examples of important linear classifiers. We encounter

- Linear discriminant analysis (LDA).
- Logistic regression.
- Support vector machines.

Tree based methods is a fourth method that relies on locally linear classifiers.

Regression

If the y variable is continuous we usually talk about regression. You should all know the linear regression model

$$Y = X^T\beta + \beta_0 + \epsilon$$

where ϵ and X are independent, $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$.

A predictor of Y given x is then a function $f : \mathbb{R}^p \to \mathbb{R}$. In the linear regression model above

$$f(x) = E(Y|X = x) = x^T\beta + \beta_0$$

is a natural choice of linear predictor.

Probability Models

Let (X, Y) be a random variable with values in $\mathbb{R}^p \times E$ and decompose its distribution P into the conditional distribution P_x of Y given $X = x$ and the marginal distribution P_1 of X. This means

$$\Pr(X \in A, Y \in B) = \int_A P_x(B) P_1(\mathrm{d}x).$$

If the joint distribution has density $g(x, y)$ w.r.t. $\nu \otimes \mu$ the marginal distribution has density

$$g_1(x) = \int g(x, y) \mu(\mathrm{d}y)$$

w.r.t. ν, the conditional distribution has density

$$g(y|x) = \frac{g(x, y)}{g_1(x)},$$

w.r.t. μ and we have Bayes formula $g(x, y) = g(y|x) g_1(x)$.

Prediction error and the Bayes classifier

If $F = \{1, \ldots, K\}$ is finite and $f : \mathbb{R}^p \to E$ is any classifier we define the expected prediction error as

$$\mathrm{EPE}(f) = P(f(X) \neq Y) = 1 - E(P_X(f(X)))$$

as the proportion of misclassifications or misclassification rate.

The Bayes classifier is the classifier that minimizes the expected prediction error and is given by

$$f_B(x) = \mathrm{argmax}_k P_x(k) = \mathrm{argmax}_k P(Y = k \mid X = x)$$

The Bayes rate

$$\mathrm{EPE}(f_B) = 1 - E(\max_k P_X(k))$$

is the expected prediction error for the Bayes classifier.

Figure 2.5 – The Bayes Classifier

The example data used for nearest neighbor are simulated and the Bayes classifier can be calculated exactly.

It can be computed using Bayes formula for $k = 0, 1$

$$\Pr(Y = k|X = x) = \frac{\pi_k g_k(x)}{\pi_0 g_0(x) + \pi_1 g_1(x)}$$

and the argmax is found to be

$$g(x) = \text{argmax}_{k=0,1} \pi_k g_k(x).$$

In the example g_0 and g_1 are mixtures of 10 Gaussian distributions.

Lecture exercise 1

Let $Y \in \{0, 1\}$ with $P(Y = 0) = 0.5$ and

$$X|(Y = k) \sim N(3k, 1).$$

Consider classifiers of the form $f_t(x) = 1(x \geq t)$ for $t \in \mathbb{R}$.

1. Write an R program that simulates a data set under the model described above with $N = 100$ independent observations (x_i, y_i). Compute and plot the empirical expected prediction error

$$t \to \frac{1}{N} \sum_{i=1} 1(f_t(x_i) \neq y_i)$$

   as a function of t.

2. Find and plot the theoretical expected prediction error

$$t \to \mathrm{EPE}(f_t)$$

   and find the Bayes classifier and Bayes rate.

Lecture exercise 2

With the same setup as above:

1. How to estimate the Bayes classifier from the data set?
2. Show that
$$P(Y = 1 \mid X = x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$
for some parameters $\alpha, \beta \in \mathbb{R}$, and show how this can be used to estimate the Bayes classifier.
3. What happens if
$$X|(Y = k) \sim N(3k, 1 + 24k).$$

Statistical Decision Theory

Question: How do we make optimal decisions of action/prediction under uncertainty?

We need to

- decide how we measure the quality of the decision – loss functions,
- decide how we model the uncertainty – probability measures,
- decide how we weigh together the losses.

Loss Functions

A loss function in the framework of ordinary regression analysis is a function $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$.

A predictor is a function $f : \mathbb{R}^p \to \mathbb{R}$. If $(x, y) \in \mathbb{R}^p \times \mathbb{R}$ the quality of predicting y as $f(x)$ is measured by the loss

$$L(y, f(x)).$$

Large values are bad! Examples where $L(y, \hat{y}) = V(y - \hat{y})$:

- The squared error loss; $V(t) = t^2$.
- The absolute value loss; $V(t) = |t|$.
- Huber for $c > 0$; $V(t) = t^2 1(|t| \leq c) + (2c|t| - c^2)1(|t| > c)$.
- The ϵ-insensitive loss; $V(t) = (|t| - \epsilon)1(|t| > \epsilon)$.

Weighing the Loss

If L is a loss function, (X, Y) a random variable and $f : \mathbb{R}^p \to \mathbb{R}$ a predictor then $L(Y, f(X))$ has a probability distribution on $[0, \infty)$.

The typical single number summary is the expected prediction error:

$$\mathrm{EPE}(f) = E(L(Y, f(X))).$$

Take Home Message: The theory depends upon choices of e.g. loss function, which often represent mathematically convenient surrogates.

Optimality is not an unconditional quality – a predictor can only be optimal given the choice of loss function, probability model and how the losses are weighed together.

Optimal Prediction

We find that

$$\begin{aligned} \mathrm{EPE}(f) &= \int L(y, f(x)) P(\mathrm{d}x, \mathrm{d}y) \\ &= \int \underbrace{\int L(y, f(x)) P_x(\mathrm{d}y)}_{E(L(Y,f(x))|X=x)} P_1(\mathrm{d}x). \end{aligned}$$

This quantity is minimized by minimizing the expected loss conditionally on $X = x$,

$$f(x) = \operatorname*{argmin}_{\hat{y}} E(L(Y, \hat{y})|X = x).$$

- Squared error loss; $L(y, \hat{y}) = (y - \hat{y})^2$

$$f(x) = E(Y|X = x)$$

- Absolute value loss; $L(y, \hat{y}) = |y - \hat{y}|$

$$f(x) = \text{median}(Y|X = x)$$

0-1 loss and the Bayes classifier

The 0-1 loss function is $L(k, l) = 1(k \neq l)$ is very popular with

$$E(L(Y, f(x))|X = x) = 1 - P_x(f(x)).$$

The optimal classifier with the 0-1 loss is the Bayes classifier already introduced and given by

$$f_B(x) = \text{argmax}_k P_x(k)$$

Linear Regression

For (X, Y) a pair of random variables with values in $\mathbb{R}^p \times \mathbb{R}$ we assume that

$$E(Y|X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j = (1, X^T)\beta$$

with $\beta \in \mathbb{R}^{p+1}$.

This "model" of the conditional expectation is linear in the parameters.

The predictor function for a given β is

$$f_\beta(x) = (1, x^T)\beta.$$

Least Squares

With $\mathbf{X}$ the $N \times (p+1)$ data matrix, including the column $\mathbf{1}$, the predicted values for given β are $\mathbf{X}\beta$.

The residual sum of squares is

$$\text{RSS}(\beta) = \sum_{i=1}^{N} (y_i - (1, x_i^T)\beta)^2 = ||\mathbf{y} - \mathbf{X}\beta||^2.$$

The least squares estimate of β is

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \text{RSS}(\beta).$$

Figure 3.1 – Geometry

The linear regression seeks a p-dimensional, affine representation – a hyperplane – of the $p + 1$-dimensional variable (X, Y).

The direction of the Y-variable plays a distinctive role – the error of the approximating hyperplane is measured parallel to this axis.

The Solution – the Calculus Way

Since $\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$

$$D_\beta \text{RSS}(\beta) = -2(\mathbf{y} - \mathbf{X}\beta)^T\mathbf{X}$$

The derivative is a $1 \times p$ dimensional matrix – a row vector. The gradient is $\nabla_\beta \text{RSS}(\beta) = D_\beta \text{RSS}(\beta)^T$.

$$D^2_\beta \text{RSS}(\beta) = 2\mathbf{X}^T\mathbf{X}.$$

If $\mathbf{X}$ has rank $p+1$, $D^2_\beta \text{RSS}(\beta)$ is (globally) positive definite and there is a unique minimizer found by solving $D_\beta \text{RSS}(\beta) = 0$. The solution is

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

Distributional Results – Conditionally on $\mathbf{X}$

$$\epsilon_i = Y_i - (1, X_i)^T \beta$$

Assumption 1: $\epsilon_1, \ldots, \epsilon_N$ are, conditionally on $X_1, \ldots, X_N$, uncorrelated with mean value 0 and same variance σ^2.

$$\hat{\sigma}^2 = \frac{1}{N-p-1} \sum_{i=1}^{N} (Y_i - \mathbf{X}\hat{\beta})^2 = \frac{1}{N-p-1} ||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2 = \frac{\text{RSS}(\hat{\beta})}{N-p-1}$$

Then $V(\mathbf{Y}|\mathbf{X}) = \sigma^2 I_N$

$$\begin{aligned} E(\hat{\beta}|\mathbf{X}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta \\ V(\hat{\beta}|\mathbf{X}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2 I_N \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\ E(\hat{\sigma}^2|\mathbf{X}) &= \sigma^2 \end{aligned}$$

Distributional Results – Conditionally on $\mathbf{X}$

Assumption 2: $\epsilon_1, \ldots, \epsilon_N$ conditionally on $X_1, \ldots, X_N$ are i.i.d. $N(0, \sigma^2)$.

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

$$(N - p - 1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p-1}.$$

The standardized Z-score

$$Z_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t_{N-p-1}.$$

Or more generally for any $a \in \mathbb{R}^{p+1}$

$$\frac{a^T\hat{\beta} - a^T\beta}{\hat{\sigma}\sqrt{a^T(\mathbf{X}^T\mathbf{X})^{-1}a}} \sim t_{N-p-1}.$$

Gauss-Markov's Theorem

Consider linear estimators only

$$\tilde{\beta} = C^T \mathbf{Y}$$

for some $N \times p$ matrix C requiring that $\beta = C^T \mathbf{X} \beta$ for all β.

Theorem

Under Assumption 1 the least squares estimator of β has minimal variance among all linear, unbiased estimators of β.

This means that for any $a \in \mathbb{R}^p$, $a^T \hat{\beta}$ has minimal variance among all estimators of $a^T \beta$ of the form $a^T \tilde{\beta}$ where $\tilde{\beta}$ is a linear, unbiased estimator.

Biased Estimators

The mean squared error is

$$\mathrm{MSE}_\beta(\tilde{\beta}) = E_\beta(||\tilde{\beta} - \beta||^2).$$

By Gauss-Markov's Theorem $\hat{\beta}$ is optimal for all β among the linear, unbiased estimators.

Allowing for biased – possibly linear – estimators we can achieve improvements of the MSE for some β – perhaps at the expense of some other β.

The Stein estimator is a non-linear, biased estimator, which under Assumption 2 has uniformly smaller MSE than $\hat{\beta}$ whenever $p \geq 3$.

Best Subset

For each $k \in \{0, \ldots, p\}$ there are

$$\binom{p}{k}$$

different models with k predictors excluding the intercept, and $p - k$ parameters $= 0$.

There are in total

$$\sum_{k=0}^{p} \binom{p}{k} = 2^p$$

different models. For the prostate dataset with $2^8 = 256$ possible models we can go through all models in a split second. With $2^{40} = 1.099.511.627.776$ we approach the boundary.

Subset Selection – A Constrained Optimization Problem

Let L_r^k for $r = 1, \ldots, \binom{p}{k}$ denote all k-dimensional subspaces of the form

$$L_r^k = \{\beta \mid p - k \text{ coordinates in } \beta = 0\}.$$

$$\hat{\beta}^k = \operatorname*{argmin}_{\beta \in \cup_r L_r^k} \mathrm{RSS}(\beta)$$

The set $\cup_r L_r^k$ is not convex – local optimality does not imply global optimality.

We can essentially only solve this problem by solving all the $\binom{p}{k}$ subproblems, which are convex optimization problems.

Conclusion: Subset selection scales computationally badly with the dimension p. Branch-and-bound algorithms can help a little ...

Figure 3.5 – Best Subset Selection

The residual sum of squares $\text{RSS}(\hat{\beta}^k)$ is a monotonely decreasing function in k.

The selected models are in general not nested.

One can not use $\text{RSS}(\hat{\beta}^k)$ to select the appropriate subset size only the best model of subset size k for each k.

Model selection criterias such as AIC and Cross-Validation can be used – these are major topics later in the course.

Test Based Selection

Set

$$\hat{\beta}^{k,r} = \operatorname*{argmin}_{\beta \in L_r^k} \mathrm{RSS}(\beta)$$

and fix $L_s^l \subseteq L_r^k$ with $l < k$.

$$F = \frac{(N-k)[\mathrm{RSS}(\hat{\beta}^{l,s}) - \mathrm{RSS}(\hat{\beta}^{k,r})]}{(k-l)\mathrm{RSS}(\hat{\beta}^{k,r})}$$

follows under Assumption 2 an F-distribution with $(k-l, N-k)$ degrees of freedom if $\beta \in L_s^l$.

L_r^k is preferred over L_s^l if $\Pr(. > F) \leq 0.05$, say – the deviation from L_s^l is unlikely to be explained by randomness alone.

Take home message: Test statistics are useful for quantifying if a simple model is inadequate compared to a complex model, but not for general model searching and selection strategies.