Machine Learning
Feature Selection

Prof. Matthias Hein

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

Lecture 16, 08.01.2014

Roadmap of the lecture

What is feature selection:

- Selection of a subset of a given feature set
- **not** related to feature construction

Motivation for feature selection:

- **Interpretation**
  - not only good prediction performance but also the question of interpretation e.g. which genes are relevant for cancer ?
  - feature selection is related to evaluation of causal effects.
- **Curse of dimensionality**
  - the smaller the dimension, the faster one can learn the dependency between features and classifier $\Longrightarrow$ better generalization

Roadmap for today

Feature selection - Theory

- The Bayes error and the feature subset selection problem
- Bayes error as a criterion for non-consistent methods
- Definitions of relevance and irrelevance of features
- Dependence Measures versus the Bayes error

Roadmap for today

Feature selection - Theory

- The Bayes error and the feature subset selection problem
- Bayes error as a criterion for non-consistent methods
- Definitions of relevance and irrelevance of features
- Dependence Measures versus the Bayes error

Feature selection is a **hard** problem !

Definition

What is feature selection:

Definition

Given a set of different features $X = \{X_1, \ldots, X_d\}$ the goal of **feature subset selection** is to extract a subset $X' = \{X_{\mu_1}, X_{\mu_2}, \ldots, X_{\mu_k}\}$ of features so that $k \ll d$ and either

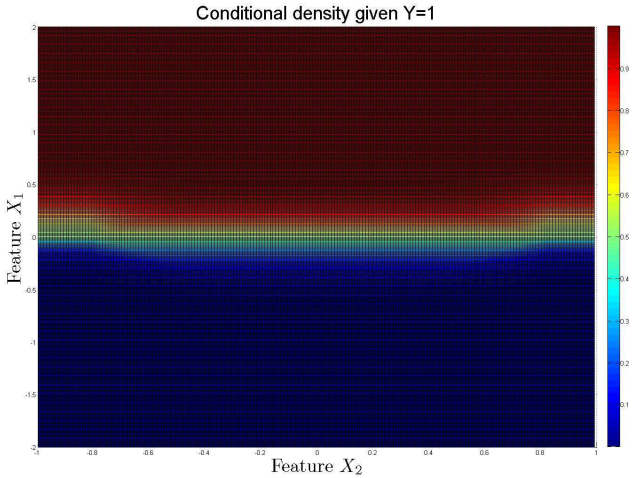
- the set of features X' is sufficient to get (almost) the same Bayes error as with the set of features X,
- the set of features X' reveal information about the target variable.

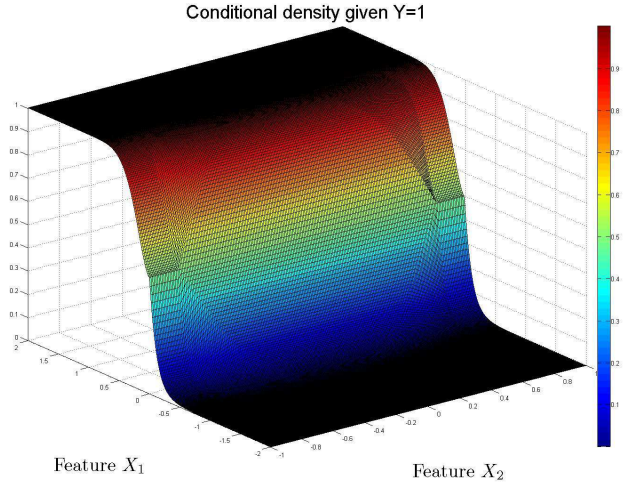
$\Longrightarrow$ Definition is quite sloppy in the criterion !

Bayes error vs Information

The two goals are not equivalent !

Density profile of the conditional density $p(x_1, x_2|Y = 1)$





- Bayes classifier: $f(x_1, x_2) = 1_{x_1 \geq 0}$ - independent of X_2
- **but** X_2 contains information about Y, mutual information $I(\{1, 2\}; Y) = 1.04$ and $I(\{1\}; Y) = 0.99$.

Bayes error

Notation: $R^*(S)$ is Bayes error for a subset of features $S \subset \{X_1, \ldots, X_d\}$.

Proposition

The Bayes risk R^* satisfies for any measurable mapping $\phi : \mathcal{X} \to \mathcal{Z}$,

$$R^*_{\mathcal{X}} \leq R^*_{\mathcal{Z}}.$$

In particular, we have for a feature subset $S \subset X = \{X_1, \ldots, X_d\}$,

$$R^*(S) \geq R^*(X).$$

Note: selection of k features corresponds to a projection $\phi : \mathbb{R}^d \to \mathbb{R}^k$,

$$\phi : (X_1, \ldots, X_d) \to (X_{\nu_1}, \ldots, X_{\nu_k}) \quad \text{with} \quad \nu_i \in \{1, \ldots, d\}.$$

One can never gain information by discarding features !

Theoretical optimum for feature selection

Definition

The **Bayes optimal feature subset** S is the smallest set of features, such that $R^*(S) = R^*(X)$.

Theoretical optimum for feature selection

Definition

The **Bayes optimal feature subset** S is the smallest set of features, such that $R^*(S) = R^*(X)$.

Why should we be interested in feature selection if it can never improve the Bayes error ?

Bayes error: theoretical quantity which gives us the best possible error **but**

- not any algorithm can find the Bayes classifier even in the limit of infinite samples. Adding features can even degrade their performance (an example for the 3-NN classifier will be given),
- the curse of dimensionality: despite the Bayes error might be smaller for a large number of features, we will never see it since we need an enormous amount of data to learn it.

Combinatorics

How many feature subsets are there ?

Given d features, there are $2^d = \sum_{k=0}^{d} \binom{d}{k}$ possible subsets of the d features.

Do we have to test **all** possibilities or is there a kind of ordering ?

Unfortunately **one has to test all** !

Result of Cover and Campenhout(1977)

Theorem

Let $S_1, S_2, \ldots, S_{2^d}$ be an ordering of the 2^d subsets of $\{1, \ldots, d\}$, satisfying the consistency property $i < j$ if $A_i \subset A_j$. Then, there exists a distribution of random variables $(X, Y) = (X_1, \ldots, X_d, Y)$ such that

$$R^*(S_1) > R^*(S_2) > \ldots > R^*(S_{2^d}).$$

Exhaustive search necessary in the worst case !

Sequential Selection

Can we select sequentially ?

No ! even if features X_i are conditionally independent given Y.

Result of Toussaint(1971):

Theorem

There exist binary-valued random variables $X_1, X_2, X_3, Y \in \{0,1\}$ *such that* X_1, X_2, X_3 *are conditionally independent given* Y *and*

$$R^*(\{2,3\}) < R^*(\{1,3\}) < R^*(\{1,2\}) < R^*(\{1\}) < R^*(\{2\}) < R^*(\{3\}).$$

Sequential selection: first feature 1 and then 2

Optimal two features: features 2 and 3 - the worst two single features

Criterion

Is the Bayes error the right criterion ?

- several classifiers converge to the Bayes classifier
- other do not. Example: 3-nearest neighbor classifier

Asymptotic error:

$$R_{\text{3-NN}} = \mathbb{E}_X[\eta(X)(1-\eta(X))\big(1+4\eta(X)(1-\eta(X))\big)].$$

Devroye et al. construct a probability measure on $[0,1]^2$ such that

$$R_{\text{3-NN}}(\{1,2\}) > R_{\text{3-NN}}(\{2\}).$$

Adding features can harm non-Bayes consistent classifiers !
Feature selection should be classifier dependent !

Relevant features

Until now: concentration on the **asymptotic performance** (Bayes error) of the classifier.

But which features contain "**relevant**" information about the target ?
$\Longrightarrow$ which feature "**influences**" the decision ?

How to define the notion of a relevant feature ?

Definition

Let $S_i = X \backslash \{X_i\}$ be the set of features with feature X_i removed. A feature X_i is **strongly relevant** if and only if Y is not conditionally independent of X_i given S_i, that is there exist x_i, y_i, s_i with $p(X_i = x_i, S_i = s_i) \neq 0$ such that

$$\mathrm{P}(Y|X_i = x_i, S_i = s_i) \neq \mathrm{P}(Y|S_i = s_i).$$

Definition

A feature X_i is **weakly relevant** if and only if it is not strongly relevant and there exists a subset $S_i' \subset S_i$ for which there exists some x_i, y and s_i' with $p(X_i = x_i, S_i' = s_i') > 0$ such that

$$\mathrm{P}(Y = y|X_i = x, S_i' = s_i') \neq \mathrm{P}(Y = y|S_i' = s_i').$$

Definition

A feature is **irrelevant** if it is neither weakly nor strongly relevant.

Illustration of relevant features

Problem:

- Five binary features $X_1, \ldots, X_5$ with values in $\{-1, 1\}$.
- We have $X_4 = -X_2$ and $X_5 = -X_3$. The eight instances of features are equally probable.
- Deterministic target is given as $Y = X_1 X_2$.

Relevance of features:

- X_1 is strongly relevant since clearly $\mathrm{P}(Y|X_1, X_2, X_3, X_4, X_5) \neq \mathrm{P}(Y|X_2, X_3, X_4, X_5)$.
- X_2 and X_4 are weakly relevant. These two features are **redundant**. Knowledge about one feature determines the other. Nevertheless, we have $\mathrm{P}(Y|X_1, X_2) \neq \mathrm{P}(Y|X_1)$.
- X_3 and X_5 are clearly irrelevant. They give no knowledge at all about the target variable Y.

Relevant features II

- Relevant features carry information about the target variable but need not be contained in the Bayes optimal feature subset.
- definition of relevant feature also not fully satisfactory.

Feature selection from the perspective of relevancy:

- **discard irrelevant features** (**But:** having a linear classifier $f(x) = \langle w, x \rangle$ and adding the irrelevant feature $X_i = 1$ changes the model into $f(x) = \langle w, x \rangle + c$ - larger capacity.)
- **keep all strongly relevant features** - all contain information about the target variable.
- **keep a minimal subset of all weakly relevant features** - eliminate all redundancy in the feature subset (difficult $\Longrightarrow$ let $X_d = f(X_1, \ldots, X_d)$ then given that f is sufficiently complicated large number of samples needed. In practice, features are often redundant.)

Dependence Measures

How can we measure the relevance of a feature (subset) ?

measure "distance" of the two probability measures: $p(S, y)$ and $p(S)p(y)$ with $S = \{X_{\nu_1}, \ldots, X_{\nu_k}\}$.

Distance metrics between probability measures:

- Hellinger Distance $d^2(P, Q) = \int_{\mathbb{R}^d} (\sqrt{p(x)} - \sqrt{q(x)})^2 \, dx$,
- Total variation $d(P, Q) = \int_{\mathbb{R}^d} |p(x) - q(x)| \, dx$,
- χ^2-distance $d^2(P, Q) = \int_{\mathbb{R}^d} \frac{(p(x)-q(x))^2}{p(x)+q(x)} \, dx$

$\Longrightarrow$ all are metrics (defined for all probability measures)

- $d(P, Q) \geq 0$ (non-negativity),
- $d(P, Q) = 0$ if and only if $P = Q$,
- $d(P, Q) = d(Q, P)$ (symmetry),
- $d(P, Q) \leq d(P, R) + d(R, Q)$ (triangle inequality),

Dependence measures from information theory

Definition

The **entropy** $H(X)$ of a random variable X is defined as

$$H(X) = -\int_{\mathcal{X}} p(x) \log_2 \left(p(x)\right) dx.$$

The **conditional entropy** $H(X_1|X_2)$ of a random variable X_1 given X_2 is defined as

$$H(X_1|X_2) = -\int_{\mathcal{X}_1} \int_{\mathcal{X}_2} p(x_1, x_2) \log_2 \left(p(x_1|x_2)\right) dx_1\, dx_2$$

- the **entropy** $H(X)$ measures the uncertainty of X: $H(X) = 0 \iff X$ deterministic, $H(X)$ is maximal for the uniform distribution given that $\mathcal{X}$ is compact,
- the **conditional entropy** measures the uncertainty of X_1 given X_2.

Dependence measures from information theory II

Definition

The **mutual information** $I(X_1; X_2)$ of two random variables X_1, X_2 is defined as

$$I(X_1; X_2) = \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} p(x_1, x_2) \log_2 \left(\frac{p(x_1, x_2)}{p(x_1)p(x_2)} \right) dx_1 \, dx_2.$$

The **conditional mutual information** $I(X_1; X_2 \,|\, X_3)$ of two random variables X_1, X_2 given X_3 is defined as

$$I(X_1; X_2 \,|\, X_3) = \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} p(x_1, x_2, x_3) \log_2 \left(\frac{p(x_1, x_2|x_3)}{p(x_1|x_3)p(x_2|x_3)} \right) dx_1 \, dx_2 \, dx_3$$

- the **mutual information** measures the dependence of X_1 and X_2,
- the **conditional mutual information** measures the conditional dependence of X_1 and X_2 given X_3.

Remarks on Definition

- In information theory one uses the logarithm with basis 2 with units "bits". Another basis of the logarithm results in a multiplicative factor,

$$\log_a(x) = \log_b(x)\ \log_a(b).$$

- For random variables on discrete sets $\mathcal{X} = \{x_1, \ldots, x_l\}$, $\mathcal{Y} = \{y_1, \ldots, y_m\}$ and $\mathcal{Z} = \{z_1, \ldots, z_n\}$ one replaces the integrals with sums:

$$H(X) = -\sum_{i=1}^{l} \mathrm{P}(x_i) \log_2 \left(\mathrm{P}(x_i)\right),$$

$$H(X|Y) = -\sum_{i=1}^{l} \sum_{j=1}^{m} \mathrm{P}(x_i, y_j) \log_2 \left(\mathrm{P}(x_i|y_j)\right),$$

$$I(X;Y) = \sum_{i=1}^{l} \sum_{j=1}^{m} \mathrm{P}(x_i, y_j) \log_2 \left(\frac{\mathrm{P}(x_i, y_j)}{\mathrm{P}(x_i)p(y_j)}\right),$$

$$I(X;Y\,|\,Z) = \sum_{i=1}^{l} \sum_{j=1}^{m} \sum_{k=1}^{n} \mathrm{P}(x_i, y_j, z_k) \log_2 \left(\frac{\mathrm{P}(x_i, y_j|z_k)}{\mathrm{P}(x_i|z_k)p(y_j|z_k)}\right).$$

Properties of these dependence measures

- $H(Y|X) = H(Y,X) - H(X)$,
- $I(X;Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$,
- $I(X;Y) = I(Y;X)$ and $I(X;Y) \geq 0$,
- $I(X;Y) = 0$ if and only if X is independent of Y,
- $I(X;Y \mid Z) = 0$ if X and Y are conditionally independent given Z.

Relation to distance metrics between probability measures:

Definition

The **f-divergence** with respect to a convex function f is defined as,

$$S_f(P,Q) = \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x)dx.$$

KL-div.: $f(t) = -\log_2(t)$, Hellinger-dist.: $f(t) = (\sqrt{t}-1)^2$, TV: $f(t) = |1-t|$.

Use of mutual information for feature selection

Maximizing dependence between feature subset and target:
Some authors define feature selection as

$$\arg\max_{|S|\leq k} I(S;Y).$$

Motivation: find all relevant features - find the maximally informative subset of features

Note: As the Bayes error monotonically decreases, the mutual information monotonically increases with the number of features.

How is this criterion related to our original one ?

$$\arg\min_{|S|\leq k} R^*(S).$$

Mutual information versus the Bayes error

Proposition

Let R^* be the Bayes risk of the zero-one loss of the data generating probability measure P on $\mathcal{X} \times \mathcal{Y}$. Then,

$$\frac{1}{2}H(Y|X) - c \quad \leq \quad R^* \quad \leq \quad \frac{1}{2}H(Y|X),$$

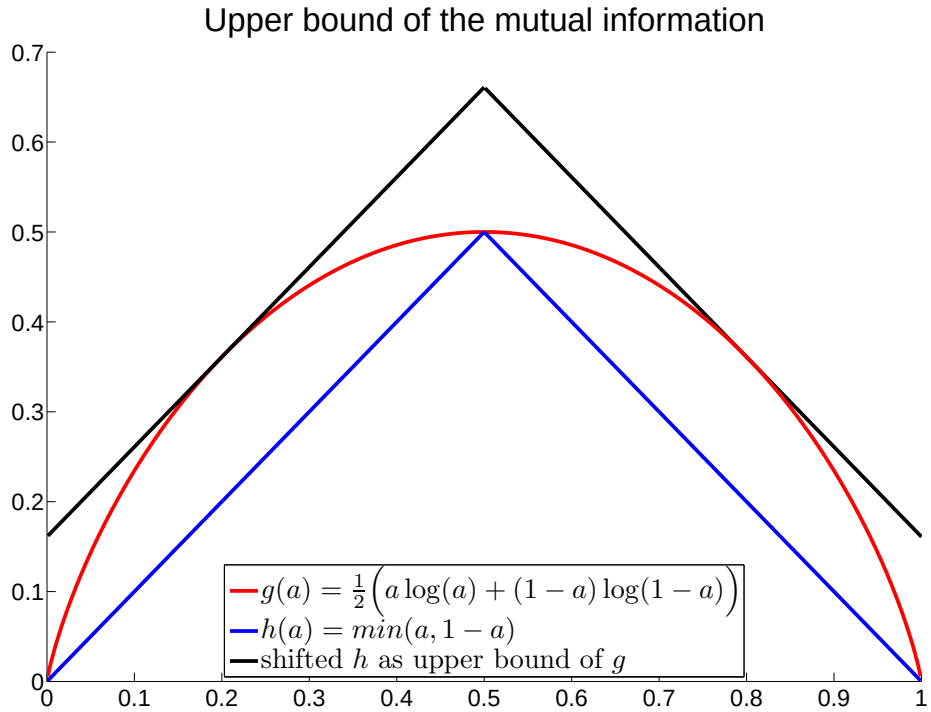
where $(X, Y) \sim \mathrm{P}$ and $c = -\frac{1}{2}\left(\frac{1}{5}\log\left(\frac{1}{5}\right) + \frac{4}{5}\log\left(\frac{4}{5}\right)\right) - \frac{1}{5} \approx 0.161$.
Expressing the conditional entropy in terms of the mutual information one obtains,

$$\frac{1}{2}[H(Y) - I(Y;X)] - c \quad \leq \quad R^* \quad \leq \quad \frac{1}{2}[H(Y) - I(Y;X)].$$

Maximizing mutual information is minimizing an upper bound on the Bayes error

Proof of the proposition

Proof by picture: function $g(a) = -\frac{1}{2}(a \log_2(a) + (1-a)\log_2(1-a))$ (red) versus $h(a) = \min\{a, 1-a\}$ (blue).



$$H(Y|X) = \int_{\mathcal{X}} \Big[-\mathrm{P}(+|x)\log_2\big(p(+|x)\big) - (1-\mathrm{P}(+|x))\log_2\big(1-\mathrm{P}(+|x)\big)\Big] p(x)\,dx,$$

$$R^* = \int_{\mathcal{X}} \min\big(\mathrm{P}(+|x), 1-\mathrm{P}(+|x)\big) p(x) dx$$

Mutual information versus Bayes error II

Implications of this Proposition:

- The upper and lower bounds do **not** imply that the total ordering of features is preserved. Let X_1 and X_2 be two features, then it can happen that $I(Y;X_1) \leq I(Y;X_2)$ but the Bayes risk of feature X_1 is smaller than the Bayes risk of X_2. This will play a role for the feature selection methods described in the next section.

Minimal feature subsets are not the same !

A feature can add information, that is mutual information about the target increases, but the Bayes error stays the same

Lemma

There exist binary-valued random variables $X_1, X_2, Y \in \{-1, 1\}$ *such that* X_1, X_2 *are conditionally independent given* Y *and*

$$R^*(\{1\}) = R^*(\{1,2\}), \qquad but \qquad I(\{1\}; Y) < I(\{1,2\}; Y).$$

Proof: The joint distribution of X_1, X_2, Y is specified by the class conditional probabilities together with $\mathrm{P}(Y = 0) = \mathrm{P}(Y = 1)$. Straightforward calculation show that

$$\mathrm{P}(X_1 = 1|Y = 0) = 0.7, \quad \mathrm{P}(X_1 = 1|Y = 1) = 0.2,$$
$$\mathrm{P}(X_2 = 1|Y = 0) = 0.8, \quad \mathrm{P}(X_2 = 1|Y = 1) = 0.6,$$

$R^*(\{1\}) = R^*(\{1,2\}) = 0.25$ but $I(\{1\}; Y) = 0.133 < 0.151 = I(\{1,2\}; Y)$.

Feature Selection in Practice

Feature selection - Practice

- Filter methods
- Wrapper methods
- Tests for linear methods

What you should not expect

- there exists no universally best method for feature selection !

Filter methods

Filter methods:

$\Longrightarrow$ **independent** of the employed learning method.

Advantages:

- faster than corresponding wrapper method (is not always true !),
- more robust to overfitting than corresponding wrapper methods.

Disadvantages:

- The best features can be classifier dependent. Classifier independent selection is suboptimal (and therefore also more robust :)).

Ideal goal in filter methods

Optimal feature subsets: Select subset $S = \{X_{\nu_1}, \ldots, X_{\nu_k}\}$ such that

$$\max_{|S| \leq k} I(Y; S).$$

The mutual information could be replaced by Bayes error, correlation,...
Alternative: penalize weighted sum of mutual information and cardinality.

Problems:

- There are $\sum_{n=0}^{k} \binom{d}{n}$ subsets of d features of cardinality smaller or equal to k. Finding the optimal feature subset is impossible (even with branch-and-bound methods) $\Longrightarrow$ greedy methods.
- the computation of the mutual information $I(Y; S)$ requires estimation of densities of up to $k + 1$ variables $\Longrightarrow$ amount of samples required for a reasonable density estimate grows exponentially with the dimension (**curse of dimensionality**). $\Longrightarrow$ replace $I(Y; S)$ with approximations.

Simple filter methods

Selection of best individual features

Compute score for each feature $\Rightarrow$ rank features according to score.

Fisher score:

$$F(i) = \frac{(m_+^{(i)} - m_-^{(i)})^2}{\sigma_{i,+}^2 + \sigma_{i,-}^2},$$

where $m_\pm^{(i)}$ and $\sigma_{i,\pm}^2$ are means and variances of feature X_i of both classes.

The Fisher score is optimal for a certain model

- individual features are conditionally independent
- class-conditional distribution of X_i is Gaussian, where variances are equal for both classes, and class probabilities are equal.

Bayes error: $R^* = \mathrm{P}\left(U > \frac{r}{2}\right), \quad U \sim \mathcal{N}(0,1)$ and $r^2 = \sum_{i=1}^{d} \frac{(m_+^{(i)} - m_-^{(i)})^2}{\sigma_i^2}$.

Selection of best individual features - continued

- **Correlation:**

$$C(i) = \frac{\sum_{j=1}^{n}(x_j^{(i)} - m^{(i)})(y_j - \bar{y})}{(n-1)\,\sigma_i\,\sigma_y},$$

where $x_j^{(i)}$ is the i-the feature of training point j, $\bar{y}$ is the mean of all labels and σ_i, σ_y are the standard deviations of the i-th feature and the class labels.

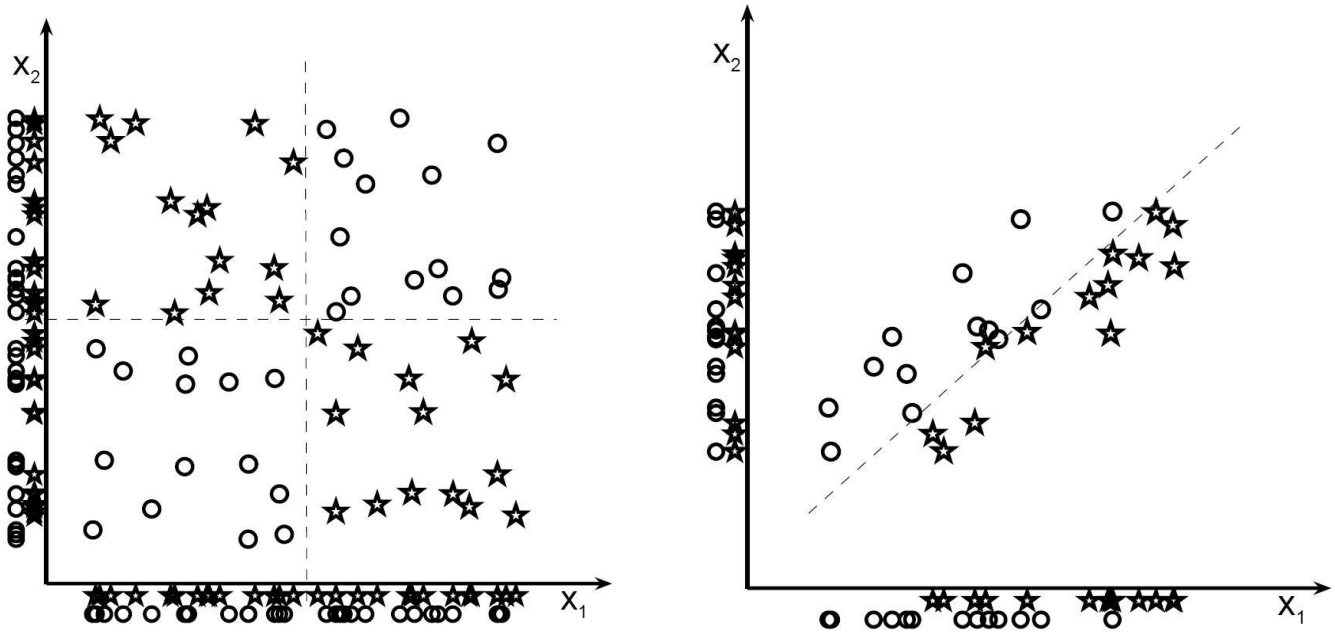
- **Mutual Information:** $I(X_i; Y)$ measures dependence of X_i and Y.

$$I(X_i, Y) = 0 \quad \Longleftrightarrow \quad X_i \text{ and } Y \text{ are independent.}$$

The (expected) correlation $C(i)$ is zero if X_i and Y are independent, but zero correlation does not mean that they are independent.
$\Longrightarrow$ better measure of independence but more difficult to compute

Problems of simple filter methods

- Correlation or Fisher scores are hardly related to the Bayes error,
- **Best individual features need not be in the best subset !**



Left: Samples of the XOR-problem. Each individual feature is usesless - Bayes error $R^*(\{1\}) = R^*(\{2\}) = 0.5$. For both features together we get $R^*(\{1,2\}) = 0$. Right: Adding the uninformative feature X_2 leads to better performance than using only feature X_1.

Filter methods - greedy methods

Sequential forward selection:

- start with empty set of features,
- add sequentially features which optimize a certain criterion.

Sequential forward using the conditional mutual information

Given X_j the conditional mutual information quantifies the gain in information about the target Y in X_i.

$$I(X_i; Y|X_j) = I(X_i, X_j; Y) - I(X_j; Y).$$

If X_i is conditionally independent of Y given X_j then $I(X_i; Y|X_j) = 0$, since $I(X_i, X_j; Y) = I(X_j; Y)$.

Filter methods - greedy methods II

Idea: Add a feature which provides the largest gain in information given the already chosen features.

- first step: initialize the feature set S with $S = \arg\max_{1 \leq i \leq d} I(X_i; Y)$,
- in the k-th step: add feature

$$X_k = \arg\max_{X_i \in X \setminus S} \min_{X_j \in S} I(X_i; Y | X_j).$$

  Add feature which maximizes the information gain given all chosen features.

Stopping criterion: pre-defined number of features or information gain drops below threshold.

Problem: The features which are chosen at some point are never discarded.

Filter methods - greedy methods III

Sequential backward selection:

- start with the full set of features,
- discard sequentially features which optimize a certain criterion.

$\Longrightarrow$ backward selection claims to detect dependencies in features more easily.

Sequential backward using the conditional mutual information

- all features are included in S,
- first step: discard feature $\arg\min_{1 \leq i \leq d} I(X_i; Y)$,
- in the k-th step: discard feature

$$X_k = \arg\min_{X_i \in S} \max_{X_j \in S} I(X_i; Y | X_j).$$

Discard feature which adds the least information given all features.

Filter methods - Branch and Bound

Branch-and-Bound

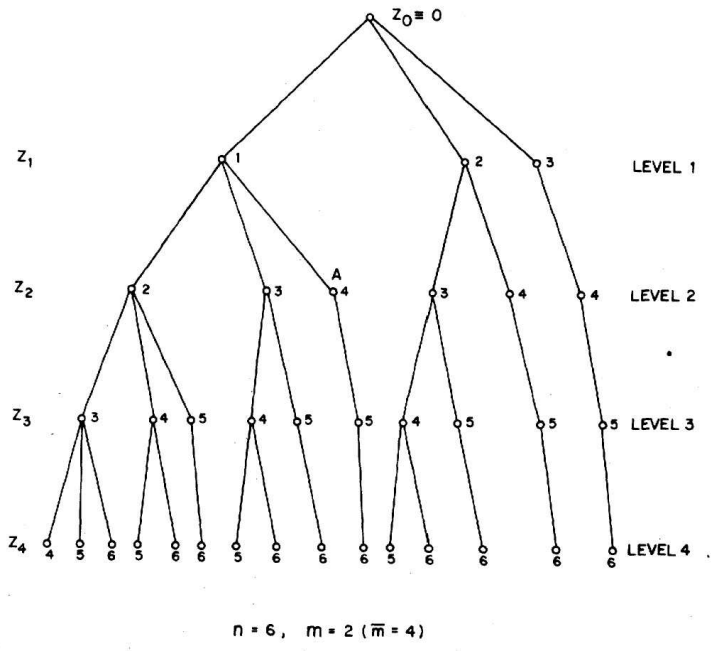
- tests all possible feature subsets for a certain criterion
- given a monotonic criterion $J(S)$, that is if $S \subset S'$, then $J(S) \geq J(S')$ the exhaustive search has not to be done completely but can be pruned.
- proposed by Narendra and Fukunaga in 1977 for classification (earlier for regression).

Possible criterion: Bayes error, mutual information.

- choose number of features k,
- create tree of all possible subsets of features which can be **discarded** $\Rightarrow \sum_{s=k}^{d} \binom{d}{s}$ such sets,
- root node is the empty set,

Filter methods - Branch and Bound II

The subset tree for the case of 6 features where 4 features are discarded.



Prune branches of the tree which have larger Bayes error (subsequent sets will even have larger Bayes error) $\Longrightarrow$ avoids exhaustive search.

Disadvantages:

- computational complexity still grows exponentially,
- Bayes error estimate uncertain - wrong branches can be discarded

$\Longrightarrow$ only possible for small feature sets of size ≤ 30.

Wrapper methods

Problem of filter methods: The evaluation criteria

- the evaluation criteria in filter methods are **independent of the learning method**.
- we are using criteria for feature selection which are only loosely connected to what we are really interested in: generalization of the learning method to future test data.

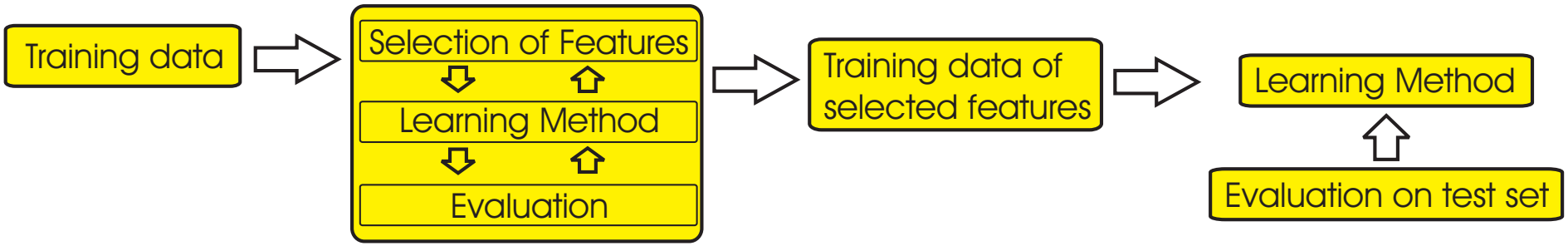
Golden principle

Always optimize the criterion which you are really interested in.

$\Longrightarrow$ Use learning method to directly evaluate the chosen feature subsets.

Wrapper methods II

General Scheme for wrapper methods:



Loop in the second step:

- select features,
- feed them into the learning method and evaluate its performance (usually cross-validation),

$\Longrightarrow$ Danger of overfitting - evaluation of final classifier on independent test set.

Wrapper methods III

Disadvantages of wrapper methods:

- For every chosen feature subset we have to train and test our learning method $\Longrightarrow$ high computational complexity.
- The feature subset selection problem is a very big model selection problem with 2^d possible models. Danger of overfitting even when one uses cross-validation for model selection. In particular, for small sample sizes one has to be very careful.

How to select features ?

Use the same techniques as in filter methods - only replace the evaluation criteria by the (cross-validation) error of the learning method.

Model specific feature selection

What is model specific feature selection ?

Assumptions about the data generating probability distribution (model)

$$\Downarrow$$

Derivation of model specific criteria e.g. the Fisher score.

Until now: main concern has been classification (but discussed methods can be immediately transferred to regression) $\Longrightarrow$ Derivation of model specific tests for the linear regression model.

The linear model

Data model: output $\mathcal{Y} = \mathbb{R}$, input $\mathcal{X} = \mathbb{R}^p$,

$$Y = \langle X, w \rangle + \varepsilon.$$

Given n samples $(X_i, Y_i)_{i=1}^n$, we have,

$$Y_i = \sum_{j=1}^{p} X_{ij} w_j + \varepsilon_j, \text{ or short } Y = Xw + \varepsilon,$$

Basic assumptions:

- error has zero mean $\mathbb{E}[\varepsilon] = 0$,
- errors of different point are uncorrelated and have same variance (homoscedastic):

$$\mathrm{Cov}(\varepsilon) = \mathbb{E}[\varepsilon \varepsilon^T] = \sigma^2 \mathbb{1}_n.$$

The linear model I

Fitting with least squares: $L(y, f(x)) = (y - f(x))^2$,

$$w_n = (X^T X)^{-1} X^T Y,$$

Gauss-Markov: w_n is unbiased estimator of w,

$$\mathbb{E}[w_n \mid T_x] = w, \text{ where } T_X = \{X_i\}_{i=1}^n.$$

Proposition

Let $\operatorname{ran}(X) = p$. *The covariance of* w_n *is given as,*

$$\operatorname{Cov}(w_n \mid T_X) = (X^T X)^{-1}\sigma^2, \quad \textit{and} \quad \hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p} X_{ij}(w_n)_j)^2,$$

is an unbiased estimator of σ^2, *that is* $\mathbb{E}(\hat{\sigma}^2 \mid T_X) = \sigma^2$.

$\Longrightarrow$ (following) results are only partially true if X has not rank p.

The linear model II

In order to design a statistical test for feature selection we need to specify the distribution of Y resp. the error ε,

$$\varepsilon \sim \mathcal{N}(0, \sigma^2).$$

When are linear transformations of Gaussian RV's independent ?

Lemma

Let $X \sim \mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$. Let $A \in \mathbb{R}^{r \times n}$ and $B \in \mathbb{R}^{s \times n}$. Then AX and BX are independent if and only if $A\Sigma B^T = 0$.

Distribution of w_n and $\hat{\sigma}^2$ in the Gaussian model

Proposition

Let $p = \operatorname{ran}(X)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ in the linear model, then

$$\hat{w}_n \sim \mathcal{N}\left(w, (X^T X)^{-1} \sigma^2\right), \quad \text{and} \quad \frac{n-p}{\sigma^2} \hat{\sigma}^2 \sim \chi^2_{N-p}.$$

Furthermore $\hat{w}_n$ and $\hat{\sigma}^2$ are independent.

χ^2-distribution and relation to Gaussian distribution

Definition

A random variable X is χ^2-distributed with parameter m or just χ^2_m distributed if it has the density,

$$p(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ K_m x^{\frac{m-2}{2}} e^{-\frac{x}{2}}, & \text{if } x > 0. \end{cases}, \quad \text{where} \quad K_m = \frac{1}{2^{\frac{m}{2}} \Gamma\left(\frac{m}{2}\right)},$$

where $\Gamma(x)$ is the Gamma-function.

Proposition

Let $Z_1, \ldots, Z_m$ be independent random variables, with $Z_i \sim \mathcal{N}(0,1)$ for $i = 1, \ldots, n$, then

$$X = \sum_{i=1}^{m} Z_i^2,$$

has a χ^2_m-distribution.

Final test statistic

Linear model:

The influence of a feature X_i is directly proportional to its weight w_i.

Definition

The **Z-score** z_j of feature X_j in the linear model is defined as

$$z_j = \frac{\hat{w}_n^{(j)}}{\sqrt{\hat{\sigma}^2 \left(X^T X\right)_{jj}^{-1}}},$$

where $\hat{w}_n^{(j)}$ is the j-th component of the weight estimate $\hat{w}_n$.

Reminder t-distribution

Definition

A random variable X is t-distributed with parameter m or just t_m distributed if it has the density,

$$p(x) = L_m \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}, \quad \text{where} \quad L_m = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{\pi\, m}\Gamma\left(\frac{m}{2}\right)}.$$

Relation of t-distribution to Gaussian and χ^2-distribution:

Proposition

Let $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2_m$ then $\frac{Z}{\sqrt{\frac{U}{m}}}$ is distributed as t_m.

Final test statistic III

Lemma

Under the null hypothesis $H_0 : w_j = 0$, z_j is distributed as t_{n-p}.

Proof: The variable $\frac{\hat{w}_n^{(j)}}{\sqrt{\sigma^2 (X^T X)^{-1}_{jj}}}$ has distribution $\mathcal{N}(0,1)$ under the null hypothesis $w_j = 0$, whereas $\hat{\sigma}^2 \frac{n-p}{\sigma^2} \sim \chi^2_{n-p}$. Thus,

$$\frac{\hat{w}_n^{(j)}}{\sqrt{\sigma^2 (X^T X)^{-1}_{jj}}} \sqrt{(n-p)\frac{\sigma^2}{(n-p)\hat{\sigma}^2}} = \frac{\hat{w}_n^{(j)}}{\sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{jj}}},$$

is distributed as t_{n-p}.

Idea: Test if the coefficient of a feature is zero (no influence on Y).

Quantiles of the t-distribution

Let X be t_m-distributed, the $1-\alpha$ quantiles for the significance level α are given in the following table.

quantile $\backslash$ m	5	10	50	100	500	1000
$\mathrm{P}(\lvert X\rvert \leq c) = 0.90$	2.015	1.813	1.676	1.660	1.648	1.646
$\mathrm{P}(\lvert X\rvert \leq c) = 0.95$	2.571	2.228	2.009	1.984	1.965	1.962
$\mathrm{P}(\lvert X\rvert \leq c) = 0.99$	4.032	3.169	2.678	2.626	2.586	2.581

Table : Given is the value $c > 0$ of the interval $[-c, c]$ which contains $1-\alpha$ of the probability mass of the t_m-distribution for different values of m and different significance levels α.

Feature Selection in the linear model

Feature selection for the linear model:

- **best subset selection:** minimizing the least squares error among all possible linear models (can be done also using branch-and-bound methods),
- **greedy forward selection:** given a linear model with k features add the feature X_i which has the highest z-score when trained with the k existing features,
- **greedy backward selection:** given a linear model with k features discard the feature X_i with the lowest z-score.

Lasso as feature selection method:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle w, \phi_i(x)\rangle)^2 + \lambda \sum_{i=1}^{D}|w_i|.$$

Trade-off between loss and used number of features (approximatively).

Comparison of different methods for a real dataset

Comparison of wrapper feature selection methods in regression:

- **Problem:** predict log-concentration of prostate-specific antigene (PSA) of for men who have prostate cancer.
- **Features:**
  1. lcavol (log cancer volume)
  2. lweight (log prostate weight)
  3. age
  4. lbph (log of the amount of benign prostatic hyperplasia)
  5. svi (seminal vesicle invasion)
  6. lcp (log of capsular penetration)
  7. gleason (gleason score)
  8. pgg45 (percent of gleason score 4 or5)
  9. intercept (the constant feature)

Comparison of different methods for a real dataset II

Comparison of wrapper feature selection methods in regression II:

- **Data:** 67 training and 30 test instances,
- **Preprocessing:** The features are centered and scaled to have unit variance (using the mean and standard deviation of the **training data**).
- **Regression method:** Least Squares linear model $f(x) = \langle x, w \rangle$
- **Evaluation method:** 5-fold cross validation error error or Z-scores.

Comparison of different methods for a real dataset III

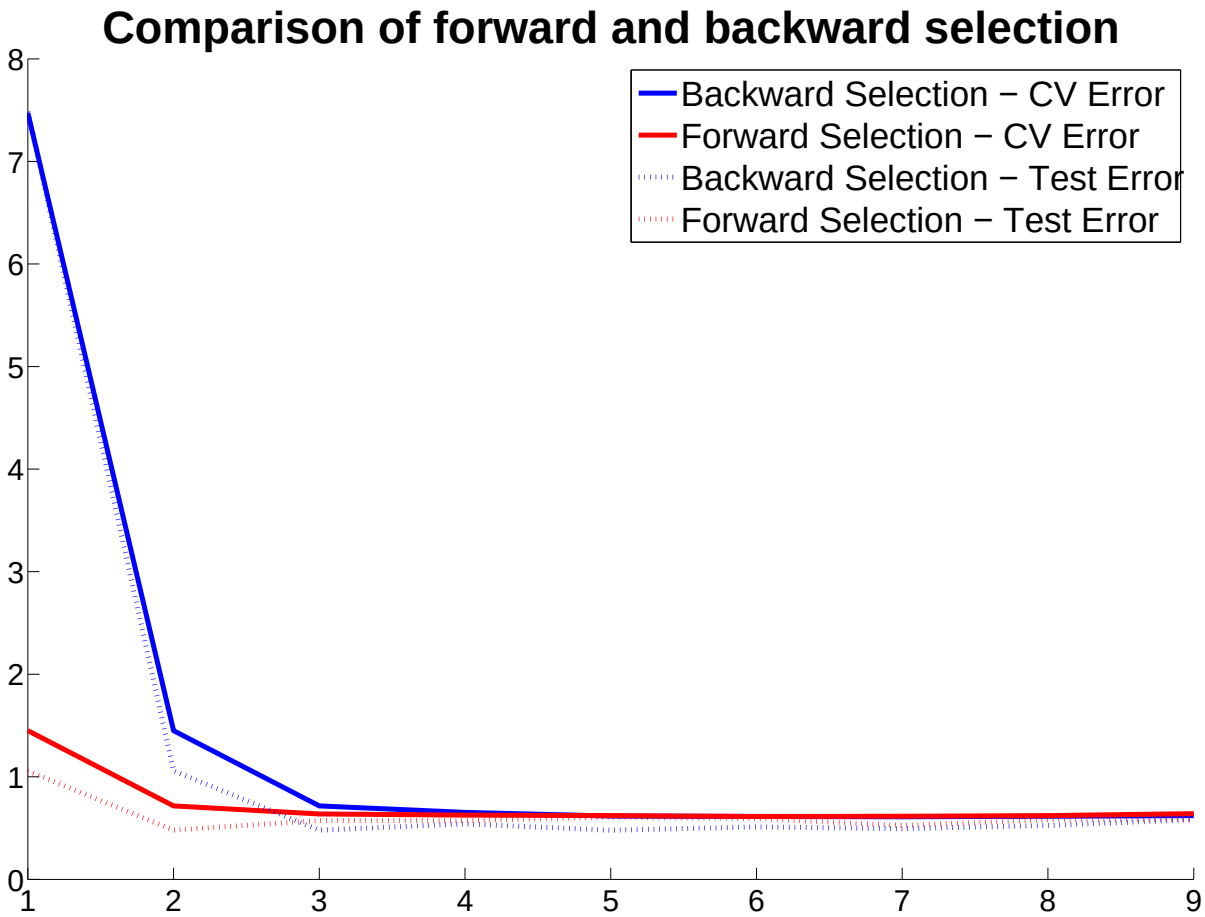
Forward and backward selection based on 5-fold cross validation:

Included (left first)	9	1	2	8	4	3	5	6	7
Forward CV error	1.45	0.72	0.64	0.63	0.62	0.61	0.62	0.62	0.64
Forward Test error	1.06	0.48	0.57	0.57	0.61	0.59	0.52	0.58	0.59

Discarded (right first)	9	1	4	5	2	3	8	6	7
Backward CV error	7.47	1.45	0.72	0.65	0.62	0.61	0.61	0.62	0.62
Backward Test error	7.48	1.06	0.48	0.54	0.48	0.51	0.50	0.53	0.58

- Results for backward selection are shown in reverse order,
- methods agree on the most important features (9=the intercept and 1=the log cancer volume) and the least valuable features (6=lcp and 7=gleason score).
- backward feature selection performs better than forward selection
- almost same performance as best feature subset selection (next frame).

Comparison of different methods for a real dataset IV



A comparison of a wrapper forward and backward selection based on 5-fold cross validation as reported in the table on the previous frame for the prediction of the PSA value.

Comparison of different methods for a real dataset V

Best feature subset selection:

Cardinality	Best CV Error	Test Error	Feature Subset
1	1.4504	1.0567	$\{9\}$
2	0.7163	0.4797	$\{9, 1\}$
3	0.6373	0.5737	$\{9, 1, 2\}$
4	0.6171	0.4785	$\{9, 1, 4, 5\}$
5	0.6131	0.5115	$\{9, 1, 4, 5, 2\}$
6	0.6099	0.4946	$\{9, 1, 4, 5, 2, 3\}$
7	0.6154	0.5254	$\{9, 1, 4, 5, 2, 3, 8\}$
8	0.6216	0.5820	$\{9, 1, 4, 5, 2, 3, 8, 6\}$
9	0.6409	0.5863	$\{9, 1, 4, 5, 2, 3, 8, 6, 7\}$

- The best subsets are almost nested (but see cardinality 3 and 4).
- Forward selection agrees with the first 3 and backward selection with the last 5 features.

Comparison of different methods for a real dataset VI

Forward and backward selection based on z-scores:

Included (left first)	9	1	2	5	4	8	6	3	7
Forward Z Score	16.62	8.70	3.58	1.99	1.99	1.19	1.73	1.49	0.15
Forward Test error	1.06	0.48	0.57	0.48	0.51	0.54	0.60	0.58	0.59

Discarded (right first)	9	1	2	5	4	8	6	3	7
Backward Z Score	16.62	8.70	3.58	1.99	1.99	1.19	1.73	1.49	0.15
Backward Test error	7.48	1.06	0.48	0.57	0.48	0.51	0.54	0.60	0.58

- absolute values of the Z-scores for the chosen/discared feature.
- forward and backward selection agree (this is generally not the case !).
- Selection based on Z-scores is better than on cross-validation.
- for $X \sim t_{67-9}$ it holds $\mathrm{P}(|X| \leq 2.002) = 0.95)$.