POLITECNICO DI MILANO — COMO CAMPUS

PATTERN ANALYSIS AND MACHINE INTELLIGENCE 2015-2016
prof. Matteo Matteucci

# Exams Summary

**with solutions**

Project Repository

| Click |
|---|

Team Members

| ID | Surname | Name |
|---|---|---|
| 10460625 | Golubeva | Svetlana |
| | | |
| | | |
| | | |

# Contents

# 1 Statistical learning (8 points)

## 1.1 ex 2015-02-09 & ex 2015-09-30

Answer the following questions:

1. Describe what are the bias, variance, and irreducible error of a model, how are they related with its complexity, how they are related to the expected prediction error, and what is the meaning of "bias-variance tradeoff"?
2. Draw a plot of (1) bias, (2) variance, (3) training error, (4) test error, and (5) irreducible error curves as a function of increasing amount of flexibility in a statistical learning method. Explain the reason of their shapes and highlight the relationships among them.

### 1.1.1 solution:

1. Lets define decomposition of expected test MSE as

$$E(y_i - \hat{f}(x_i))^2 = Var(\hat{f}(x_i)) + [Bias(\hat{f}(x_i))]^2 + Var(\epsilon)$$

Where:
$[Bias(\hat{f}(x_i))]^2$ is a squared bias and refers to the error that is introduced by approximating a real-life problem by much simpler problem (some important unobserved predictors / influences can exist, or some features can be missed during simplification). Generally, the more flexible method we use then the less bias we have.

$Var(\hat{f}(x_i))$ is a variance and refers to the amount by which $\hat{f}$ would change if we estimated it using a different data set. Generally, the more flexible method we use then the higher variance we have.

$Var(\epsilon)$ is a irreducible error of a model and means the variance of the error term. Its presence appears due to errors of measurement. In other words, it is the noise term in the true relationship that cannot fundamentally be reduced by any model. Analogous to the Bayes error rate which is the lowest possible error rate for any classifier of a random outcome.

Relation to the expected prediction error. Equation $E(y_i - \hat{f}(x_i))^2$ defines the expected test MSE (alternative name for it is Expected Prediction error, EPE). To minimize this quantity we need to find such statistical learning method, that simultaneosly achieves low bias and low variance — the meaning of "bias-variance tradeoff". In other words, it allows us to find such optimum where training and test MSE are weighted equally (test not grow, training not fall).

Relation with a model complexity. In general, the main task is to minimize the overall error and choose an optimal level of complexity of a model.

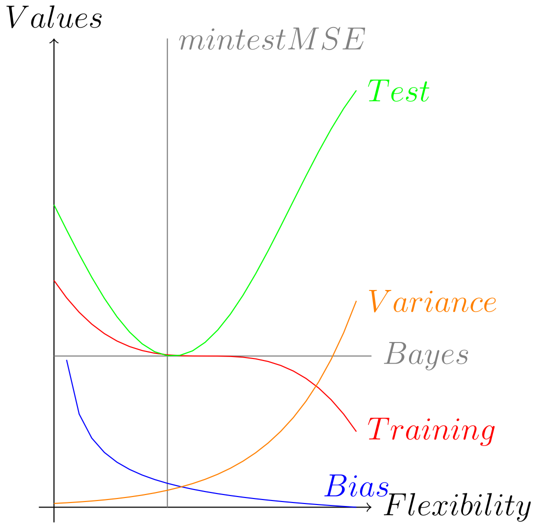
It means that for any model the level of complexity at which the increase in bias is equivalent to the reduction in variance. Mathematically:

$$\frac{dBias}{dComplexity} = -\frac{dVariance}{dComplexity}$$

If the model complexity exceeds this value, we are in effect over-fitting our model; while if our complexity falls short of the value, we are under-fitting the model.

In practice, there is no analytical way to find such point. Instead we must use an accurate measure of prediction error and explore differing levels of model complexity and then choose its level that minimizes the overall error.

2. The blue line represents the bias curve (1). It tends to decrease with level of complexity growth and intersects variance curve in point of optimal complexity of a model.
The orange line represents the variance curve (2). It tends to increase with level of complexity growth and intersects bias curve in point of optimal complexity of a model.
The red line represents the training error curve (3). It has "U"-shaped curve with saddle point in place of optimal complexity of a model.
The green line represents the test error curve (4).It has shape analoguos to bias with "fall" close to the end.
The horizontal grey line represents the Bayes (irreducible) error (5). Its is usually constant.
The vertical grey line indicates the flexibility level corresponding to the smallest test MSE.



## 1.2 ex 2015-02-23

In statistical learning theory, Test and Training set Mean Squared Errors are related by the so called Bias-Variance trade-off:

1. Write and comment the formula representing the Bias-Variance trade off for the Expected Prediction Error in Regression

The previous formula does not hold for Classification, but a useful result exists for the Classification Error Rate as well

2. Write and comment what statistical learning theory states about the minimum achievable average test error rate.

Provided the previous result for Classification, answer the following questions

3. Describe in detail how the previous result for the optimal classifier is used to derive the Logistic Regression classifier; provide a detailed description of the underlining model for Logistic Regression and derive the shape of the decision boundary for Logistic Regression.
4. Describe in detail how the previous result for the optimal classifier is used to derive Linear Discriminant Analysis; provide a detailed description of the underlining model for Linear Discriminant Analysis and derive the shape of the decision boundary for Linear Discriminant Analysis.

### 1.2.1 solution:

1. Expected Prediction Error can be expressed as

$$E(y_i - \hat{f}(x_i))^2 = Var(\hat{f}(x_i)) + [Bias(\hat{f}(x_i))]^2 + Var(\epsilon)$$

Where:
$[Bias(\hat{f}(x_i))]^2$ is a squared bias and refers to the error that is introduced by approximating a real-life problem by much simpler problem (some important unobserved predictors / influences can exist, or some features can be missed during simplification). Generally, the more flexible method we use then the less bias we have.

$Var(\hat{f}(x_i))$ is a variance and refers to the amount by which $\hat{f}$ would change if we estimated it using a different data set. Generally, the more flexible method we use then the higher variance we have.

$Var(\epsilon)$ is a irreducible error of a model and means the variance of the error term. Its presence appears due to errors of measurement. In other words, it is the noise term in the true relationship that cannot fundamentally be reduced by any model. Analogous to the Bayes error rate which is the lowest possible error rate for any classifier of a random outcome.

2. In case of Classification, we can use the Classification Error Rate (also named Bayes error rate). It is the lowest possible error rate for any classifier of a random outcome and is analogous to the irreducible error.

The Bayes error rate of the data distribution is the probability an instance is misclassified by a classifier that knows the true class probabilities given the predictors. For a multiclass classifier, the Bayes error rate may be calculated as follows:

$$p = \int_{x \in H_i} \sum_{C_i \neq C_{\text{max,x}}} P(C_i|x)p(x)\,dx$$

Where $x$ is an instance, $C_i$ is a class into which an instance is classified, $H_i$ is the area/region that a classifier function $h$ classifies as $C_i$.

The Bayes error is non-zero if the classification labels are not deterministic (there is a non-zero probability of a given instance belonging to more than one class).

3. Logistic Regression is a regression model where the dependent variable is categorical (belongs to some class).

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probit regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors.

Logistic regression can be seen as a special case of the generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of linear regression. In particular the key differences of these two models can be seen in the following two features of logistic regression. First, the conditional distribution $y \mid x$ is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary. Second, the predicted values are probabilities and are therefore restricted to (0,1) through the logistic distribution function because logistic regression predicts the probability of particular outcomes.

provide a detailed description of the underlining model for Logistic Regression

Decision boundary is given by logistic function $F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$.

4. Linear Discriminant Analysis

provide a detailed description of the underlining model for Linear Discriminant Analysis

Decision boundary is given by .

## 1.3 ex 2015-07-06

(a) Both classification and regression problems can be solved by simple but restrictive methods as well as more complex but flexible ones. Explain which ones are better:

1. when doing inference or prediction;
2. when the irreducible error is extremely high;
3. when the number of observations is very large and the number of predictors is small;
4. when the function we need to estimate is highly non-linear.

(b) What are the Bayes classifier and the Bayes error rate? Define them and explain why the latter is defined as analogous to the irreducible error.

### 1.3.1 solution:

## 1.4 ex 2015-09-14

In statistical learning theory, Test and Training set Mean Squared Errors are related by the so called Bias-Variance trade-off:

(a) Write and comment the formula representing the Bias-Variance trade off for the Expected Prediction Error in Regression

(b) The previous formula does not hold for Classification, but a useful result exists for the Classification Error Rate; write and comment what statistical learning theory states about the minimum achievable average test error rate.

(c) Describe in detail how the previous result for the optimal classifier is used to derive Linear Discriminant Analysis (LDA) classifier providing a detailed description of the underlining model and the shape of its decision boundary (derive it from the model).

(d) Train a LDA classifier using the data provided in the table and provide the classification error achieved by this classifier on the training data

| X | Class |
|---|---|
| 0 | A |
| 1 | A |
| 2 | A |
| 1 | A |
| 3 | A |
| 1 | B |
| 2 | B |
| 3 | B |
| 4 | B |
| 4 | C |
| 6 | C |
| 6 | C |
| 6 | C |

### 1.4.1 solution:

(a) The Bias-Variance trade-off for the Expected Prediction Error in Regression:

(b) The Classification Error Rate; statistical learning theory states about the minimum achievable average test error rate.

(c) Describe in detail how the previous result for the optimal classifier is used to derive Linear Discriminant Analysis (LDA) classifier providing a detailed description of the underlining model and the shape of its decision boundary (derive it from the model).

(d.1) LDA classifier training:

| X | Class | $\hat{\mu}$ | $x-\hat{\mu}$ | $(x-\hat{\mu})^2$ | $\sum(x-\hat{\mu})^2$ |
|---|---|---|---|---|---|
| 0 | A | 1.4 | -1.4 | 1.96 | 5.2 |
| 1 | A | | -0.4 | 0.16 | |
| 2 | A | | 0.6 | 0.36 | |
| 1 | A | | -0.4 | 0.16 | |
| 3 | A | | 1.6 | 2.56 | |
| 1 | B | 2.5 | -1.5 | 2.25 | 5 |
| 2 | B | | -0.5 | 0.25 | |
| 3 | B | | 0.5 | 0.25 | |
| 4 | B | | 1.5 | 2.25 | |
| 4 | C | 5.5 | -1.5 | 2.25 | 3 |
| 6 | C | | 0.5 | 0.25 | |
| 6 | C | | 0.5 | 0.25 | |
| 6 | C | | 0.5 | 0.25 | |

# of classes $K = 3$, overall # of points $n = 13$, # of points in class A is $n_1 = 5$, # of points in class B is $n_2 = 4$, # of points in class C is $n_3 = 4$;

Compute $\hat{\mu} = \frac{1}{n_k}\sum x$, where $\sum x$ is sum over all elements in the class and $n_k$ - # of elements in this class for each of three classes in given dataset:

$$\hat{\mu_1} = \frac{\sum x}{n_1} = \frac{0+1+2+1+3}{5} = \frac{7}{5} = 1.4$$

$$\hat{\mu_2} = \frac{\sum x}{n_2} = \frac{1+2+3+4}{4} = \frac{10}{4} = 2.5$$

$$\hat{\mu_3} = \frac{\sum x}{n_2} = \frac{4+6+6+6}{4} = \frac{22}{4} = 5.5$$

Compute a common variance $\hat{\sigma}^2 = \frac{1}{n-K}\sum\sum(x-\hat{\mu})^2$

$$\hat{\sigma}^2 = \frac{5.2+5+3}{13-3} = \frac{13.2}{10} = 1.32$$

Compute the class membership probabilities $\pi_k = \frac{n_k}{n}$ and its logarithm $log(\pi_k)$:

$$\pi_1 = \frac{n_1}{n} = \frac{5}{13} = 0.38, log(\pi_1) = log(\frac{5}{13}) = -0.41$$

$$\pi_2 = \frac{n_2}{n} = \frac{4}{13} = 0.31, log(\pi_2) = log(\frac{4}{13}) = -0.51$$

$$\pi_3 = \frac{n_3}{n} = \frac{4}{13} = 0.31, log(\pi_3) = log(\frac{4}{13}) = -0.51$$

The decision boundary for LDA classifier (discriminant function) is given by $\hat{\delta}_k(x) = x\frac{\hat{\mu_k}}{\hat{\sigma}^2} - \frac{\hat{\mu_k}^2}{2\hat{\sigma}^2} + log(\pi_k)$. For simplicity denote $a_k = \frac{\hat{\mu_k}}{\hat{\sigma}^2}$ and $b_k = log(\pi_k) - \frac{\hat{\mu_k}^2}{2\hat{\sigma}^2}$. So the equation takes the form $\hat{\delta}_k(x) = xa_k + b_k$

Compute $a_k = \frac{\hat{\mu_k}}{\hat{\sigma}^2}$:

$$a_1 = \frac{\hat{\mu_1}}{\hat{\sigma}^2} = \frac{1.4}{1.32^2} = \frac{1.4}{1.74} = 0.8$$

$$a_2 = \frac{\hat{\mu_2}}{\hat{\sigma}^2} = \frac{2.5}{1.32^2} = \frac{2.5}{1.74} = 1.44$$

$$a_3 = \frac{\hat{\mu_3}}{\hat{\sigma}^2} = \frac{5.5}{1.32^2} = \frac{5.5}{1.74} = 3.16$$

Compute $b_k = log(\pi_k) - \frac{\hat{\mu_k}^2}{2\hat{\sigma}^2}$:

$$b_1 = log(\pi_1) - \frac{\hat{\mu_1}^2}{2\hat{\sigma}^2} = log(\frac{5}{13}) - \frac{1.4^2}{2 * 1.32^2} = -0.41 - \frac{1.96}{3.48} = -0.41 - 0.56 = -0.97$$

$$b_2 = log(\pi_2) - \frac{\hat{\mu_2}^2}{2\hat{\sigma}^2} = log(\frac{4}{13}) - \frac{2.5^2}{2 * 1.32^2} = -0.51 - \frac{6.25}{3.48} = -0.51 - 1.8 = -2.31$$

$$b_3 = log(\pi_3) - \frac{\hat{\mu_3}^2}{2\hat{\sigma}^2} = log(\frac{4}{13}) - \frac{5.5^2}{2 * 1.32^2} = -0.51 - \frac{30.25}{3.48} = -0.51 - 8.7 = -9.21$$

We get the following equations for decision boundaries:

$$\hat{\delta_1}(x) = 0.8x - 0.97$$

$$\hat{\delta_2}(x) = 1.44x - 2.31$$

$$\hat{\delta_3}(x) = 3.16x - 9.21$$

We assign an observation X = x to the class for which $\hat{\delta}_k(x)$ is largest.

(d.2) The classification error (on the training data):

## 1.5 ex 2016-02-03

According to statistical learning theory, in regression we assume a relationship exists between an observed variable and and a dependent variable in the form

$$Y_i = f(X_i) + \epsilon_i, \epsilon_i \sim N(o, \sigma^2)$$

1. What are the *two* sources of errors we have whe estimating $f$ from data and what are these errors due to?
2. According to statistical learning theory, Test and Training Mean Squared Errors are related by Bias-Variance trade-off; write and comment the formula representing the Bias-Variance trade-off for the Expected Prediction Error in Regression.
3. The previos formula does not hold for Classification, but a useful result exists for the Classification Error Rate as well. Write and comment what statistical learnig theory states about the minimal achievable average test error rate.
4. Describe in detail how the previous result for the optimal classifier is used to derive the Logistic Regression classifier and derive the shape of the decision boundary for Logistic Regression.

### 1.5.1 solution:

## 1.6 ex 2016-02-19 Generative vs. Disciriminative models

A classical distinctions between classification models is the generative vs discriminative one. Answer the following about this distinction.

(a) What are discriminative and generative models? How do they differ? Which one should be preferred and why?

(b) Is Logistic Regression a discriminative model or a generative one? Why?

(c) Is Linear Discriminant Analysis a discriminative model or a generative one? Why?

(d) Is Support Vector Machines a discriminative model or a generative one? Why?

Let us consider the Support Vector Machine model for classification

(e) What is a Support Vector Machine? How it is defined (i.e., the optimization problem it solves) and how is it trained (i.e., the optimization problem is solved to train it)? How does the solution looks like and what this has to do with the name of the model?

(f) What is the kernel trick and how can it be applied to Support Vector Machines (i.e., what do you need to change with respect to the original algorithm)?

### 1.6.1 solution:

(a) Discriminative model (conditional model) is a model used in machine learning for modeling the dependence of an unobserved variable $y$ on an observed variable $x$. Within a probabilistic framework, this is done by modeling the conditional probability distribution $P(y|x)$, which can be used for predicting $y$ from $x$ .

Discriminative models, as opposed to generative models, do not allow one to generate samples from the joint distribution of $x$ and $y$. However, for tasks such as classification and regression that do not require the joint distribution, discriminative models can yield superior performance.

On the other hand, generative models are typically more flexible than discriminative models in expressing dependencies in complex learning tasks. In addition, most discriminative models are inherently supervised and cannot easily be extended to unsupervised learning. Application specific details ultimately dictate the suitability of selecting a discriminative versus generative model. Examples

Examples of discriminative models used in machine learning include:

Linear regression Logistic regression Support vector machines Boosting (meta-algorithm) Conditional random fields Neural networks Random Forests

Generative model is a model for randomly generating observable data values, typically given some hidden parameters. It specifies a joint probability distribution over observation and label sequences. Generative models are used in machine learning for either modeling data directly (i.e., modeling observations drawn from a probability density function), or as an intermediate step to forming a conditional probability density function. A conditional distribution can be formed from a generative model through Bayes' rule.

Generative models contrast with discriminative models, in that a generative model is a full probabilistic model of all variables, whereas a discriminative model provides a model only for the target variable(s) conditional on the observed variables. Thus a generative model can be used, for example, to simulate (i.e. generate) values of any variable in the model, whereas a discriminative model allows only sampling of the target variables conditional on the observed quantities. Despite the fact that discriminative models do not need to model the distribution of the observed variables, they cannot generally express more complex relationships between the observed and target variables. They don't necessarily perform better than generative models at classification and regression tasks. In modern applications the two classes are seen as complementary or as different views of the same procedure.[1]

Examples of generative models include:

Gaussian mixture model and other types of mixture model Hidden Markov model Probabilistic context-free grammar Naive Bayes Averaged one-dependence estimators Latent Dirichlet allocation Restricted Boltzmann machine

If the observed data are truly sampled from the generative model, then fitting the parameters of the generative model to maximize the data likelihood is a common method. However, since most statistical models are only approximations to the true distribution, if the model's application is to infer about a subset of variables conditional on known values of others, then it can be argued that the approximation makes more assumptions than are necessary to solve the problem at hand. In such cases, it can be more accurate to model the conditional density functions directly using a discriminative model (see above), although application-specific details will ultimately dictate which approach is most suitable in any particular case.

Differences:

Use case:

(b) Logistic Regression (LogR) is a discriminative model, because

(c) Linear Discriminant Analysis (LDA) is a generative model, because

(d) Support Vector Machines (SVM) is a discriminative model, because

(e) Support Vector Machine is a

How it is defined (i.e., the optimization problem it solves) and

how is it trained (i.e., the optimization problem is solved to train it)?

How does the solution looks like and what this has to do with the name of the model?

(f) The kernel trick is a

how can it be applied to Support Vector Machines (i.e., what do you need to change with respect to the original algorithm)?

## 1.7 ex 2016-07-06

According to statistical learning theory, in regression we assume a relationship exists between an observed variable and and a dependent variable in the form

$$Y_i = f(X_i) + \epsilon_i, \epsilon_i \sim N(o, \sigma^2)$$

1. What are the *two* sources of errors we have whe estimating $f$ from data and what are these errors due to?
2. According to statistical learning theory, Test and Training Mean Squared Errors are related by Bias-Variance trade-off; write and comment the formula representing the Bias-Variance trade-off for the Expected Prediction Error in Regression.
3. How is model complexity related to Bias and Variance? First provide a definition of model complexity, then according to that definition explain how bias and variance are influenced by an increased model complexity and why.
4. Provide two different examples for $f(X_i)$, describe their trade-off in terms of Bias and Variance, i.e., if one of term reduces either bias or variance with respect to the other, and explain when one should prefer one of the two models with respect to the other.

### 1.7.1 solution:

# 2 Linear regression (8 points)

## 2.1 ex 2015-02-09 & ex 2016-07-06

Given the variables x = {1, 2, 3, 4, 5, 6, 7, 8, 9} and y = {3.3, 3.6, 5.2, 5.6, 7.4, 8.3, 8.7, 9.7, 11.2}

1. Manually compute the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ of a linear model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ which fits the given data
2. What is the value of MSE calculated between the values of y and the ones returned by the $\hat{y}$ function?
3. How can we compute if the trend identified by $\hat{\beta}_1$ is significant or it is just due to spurious correlations?

To ease your computation, you can follow the following steps:

- calculate the mean $\bar{x}$ of x
- calculate the mean $\bar{y}$ of y
- calculate $x - \bar{x}$ (a vector where each value is $x_i - \bar{x}$ )
- calculate $y - \bar{y}$ (as above)
- calculate $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
- calculate $\hat{\beta}_o = \hat{y} - \hat{\beta}_1 x$

### 2.1.1 solution

1)

| x | y | $x - \bar{x}$ | $y - \bar{y}$ | $(x - \bar{x})^2$ | $(x - \bar{x})(y - \bar{y})$ | $\hat{y}$ | $(y - \hat{y})^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 3.3 | -4 | -3.7 | 16 | 14.8 | 3.02667 | 0.0747108 |
| 2 | 3.6 | -3 | -3.4 | 9 | 10.2 | 4.02 | 0.1764 |
| 3 | 5.2 | -2 | -1.8 | 4 | 3.6 | 5.01333 | 0.0348444 |
| 4 | 5.6 | -1 | -1.4 | 1 | 1.4 | 6.00667 | 0.165378 |
| 5 | 7.4 | 0 | 0.4 | 0 | 0 | 7 | 0.16 |
| 6 | 8.3 | 1 | 1.3 | 1 | 1.3 | 7.99333 | 0.0940448 |
| 7 | 8.7 | 2 | 1.7 | 4 | 3.4 | 8.98667 | 0.0821773 |
| 8 | 9.7 | 3 | 2.7 | 9 | 8.1 | 9.98 | 0.0783999 |
| 9 | 11.2 | 4 | 4.2 | 16 | 16.8 | 10.9733 | 0.0513781 |

.

The formula for calculating of the sample mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, so

$\bar{x} = \frac{1+2+3+4+5+6+7+8+9}{9} = \frac{45}{9} = 5$

$\bar{y} = \frac{3.3+3.6+5.2+5.6+7.4+8.3+8.7+9.7+11.2}{9} = \frac{63}{9} = 7$

$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 59.6$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 60$

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{59.6}{60} = 0.99(3) \approx 1$

$\hat{\beta}_o = \hat{y} - \hat{\beta}_1 x = 7 - 0.99(3) * 5 = 2.0(3) \approx 2$

2)
MSE is Mean Squared error, computes by $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$

For given data set and estimated parameters we have MSE = 0.101926, which means that we have 10% error in our prediction.

## 2.2 ex 2015-02-23

Provide detailed answers to the following

1. Let assume you have a dataset with n=1000 observations and you try to fit different models on the data:

   - A linear regression model, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$
   - The polynomial regression model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$
   - A smoothing spline (i.e. an even more flexible model than the previous two)

   For each of the three models, you calculate both training and test RSS. How would you expect the values of RSS to be (both in the training and in the test case), supposing that the true relationship between X and Y is (a) linear or (b) cubic?

2. What is the additive assumption in a linear regression model? Show how you would detect and quantify a possible interaction between the variables in a regression model and how you would model it from a statistical perspective. Finally explain, with an example, how your model would take this interaction into account (e.g. explain how, given a change in the input, the dependent variable and consequently the output change).

## 2.3 ex 2015-07-06

Given the variables $x_1 = \{14, 1, 13, 8, 11, 19, 0, 9\}$, $x_2 = \{12, 13, 7, 10, 8, 11, 16, 10\}$, and $y = \{26, 7, 24, 15, 24, 38, 2, 13\}$, manually calculate the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ of the two linear functions $\hat{y} = \beta_0 + \beta_1 x_1$ and $\hat{y} = \beta_0 + \beta_1 x_2$ which fit the given data. To ease your calculations, take the following steps:

1. calculate the mean $\bar{x}$ of x and the mean $\bar{y}$ of y
2. calculate $x - \bar{x}$ (a vector where each value is $x_i - \bar{x}$ ) and $y - \bar{y}$
3. calculate $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, then $\hat{\beta}_o = \hat{y} - \hat{\beta}_1 x$

Measure the quality of your predictions in terms of MSE, compare the results of the two linear regressions, and justify them.

### 2.3.1 solution

For the first set $x_1 = \{14, 1, 13, 8, 11, 19, 0, 9\}$ and $y = \{26, 7, 24, 15, 24, 38, 2, 13\}$:

| $x_1$ | y | $x_1 - \bar{x}$ | $y - \bar{y}$ | $(x_1 - \bar{x_1})^2$ | $(x_1 - \bar{x_1})(y - \bar{y})$ |
|---|---|---|---|---|---|
| 14 | 26 | 4.625 | 7.375 | 21.3906 | 34.1094 |
| 1 | 7 | -8.375 | -11.625 | 70.1406 | 97.3594 |
| 13 | 24 | 3.625 | 5.375 | 13.1406 | 19.4844 |
| 8 | 15 | -1.375 | -3.625 | 1.89062 | 4.98438 |
| 11 | 24 | 1.625 | 5.375 | 2.64062 | 8.73438 |
| 19 | 38 | 9.625 | 19.375 | 92.6406 | 186.484 |
| 0 | 2 | -9.375 | -16.625 | 87.8906 | 155.859 |
| 9 | 13 | -0.375 | -5.625 | 0.140625 | 2.10938 |

.

The formula for calculating of the sample mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, so

$\bar{x} = \quad = 9.375$

$\bar{y} = \quad = 18.675$

$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 509.125$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 289.875$

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \quad = 1.75636 \approx$

$\hat{\beta}_o = \hat{y} - \hat{\beta}_1 x == 2.15912 \approx$

For the second set $x_2 = \{12, 13, 7, 10, 8, 11, 16, 10\}$, and $y = \{26, 7, 24, 15, 24, 38, 2, 13\}$:

| $x_2$ | y | $x_2 - \bar{x}$ | $y - \bar{y}$ | $(x_2 - \bar{x_2})^2$ | $(x_2 - \bar{x_2})(y - \bar{y})$ |
|---|---|---|---|---|---|
| 12 | 26 | 1.125 | 7.375 | 1.26562 | 8.29688 |
| 13 | 7 | 2.125 | -11.625 | 4.51562 | -24.7031 |
| 7 | 24 | -3.875 | 5.375 | 15.0156 | -20.8281 |
| 10 | 15 | -0.875 | -3.625 | 0.765625 | 3.17188 |
| 8 | 24 | -2.875 | 5.375 | 8.26562 | -15.4531 |
| 11 | 38 | 0.125 | 19.375 | 0.015625 | 2.42188 |
| 16 | 2 | 5.125 | -16.625 | 26.2656 | -85.2031 |
| 10 | 13 | -0.875 | -5.625 | 0.765625 | 4.92188 |

.

The formula for calculating of the sample mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, so

$\bar{x} = \ = 10.875$

$\bar{y} = \ = 18.625$

$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = -127.375$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 56.875$

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \ = -2.23956 \approx$

$\hat{\beta}_o = \hat{y} - \hat{\beta}_1 x == 42.9802 \approx$

## 2.4 ex 2015-09-14

(a) What is the standard error and how is it used to calculate a confidence interval? For instance, what does it mean to have a 95% confidence interval on the parameter $\beta_1$ of a linear regression?

(b) Explain what the null hypothesis is in the context of linear regression and how it is verified.

## 2.5 ex 2015-09-30

Given the following observations x = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} and y = {12, 11.2, 9.7, 8.7, 8.3, 7.4, 5.6, 5.2, 3.6, 3.3}

1. Manually compute the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ of a linear model $\hat{y} = \beta_0 + \beta_1 x$ which fits the given data

2. What is the value of MSE calculated between the values of y and the ones returned by the $\hat{y}$ function?

3. Is the trend identified by $\hat{\beta}_1$ significant or it is just due to spurious correlations? You have to provide supporting computations and justifications for your answer.

### 2.5.1 solution

| x | y | $x - \bar{x}$ | $y - \bar{y}$ | $(x - \bar{x})^2$ | $(x - \bar{x})(y - \bar{y})$ |
|---|---|---|---|---|---|
| 1 | 12 | -4.5 | 4.5 | 20.25 | -20.25 |
| 2 | 11.2 | -3.5 | 3.7 | 12.25 | -12.95 |
| 3 | 9.7 | -2.5 | 2.2 | 6.25 | -5.5 |
| 4 | 8.7 | -1.5 | 1.2 | 2.25 | -1.8 |
| 5 | 8.3 | -0.5 | 0.8 | 0.25 | -0.4 |
| 6 | 7.4 | 0.5 | -0.1 | 0.25 | -0.05 |
| 7 | 5.6 | 1.5 | -1.9 | 2.25 | -2.85 |
| 8 | 5.2 | 2.5 | -2.3 | 6.25 | -5.75 |
| 9 | 3.6 | 3.5 | -3.9 | 12.25 | -13.65 |
| 10 | 3.3 | 4.5 | -4.2 | 20.25 | -18.9 |

.

The formula for calculating of the sample mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, so

$\bar{x} = \; = 5.5$

$\bar{y} = \; = 7.5$

$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = -82.1$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 82.5$

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \; = -0.995152 \approx$

$\hat{\beta}_o = \hat{y} - \hat{\beta}_1 x == 12.9733 \approx$

## 2.6 ex 2016-02-03

Given the following observations x = {45, 54, 41, 55, 52, 56, 49, 50, 46, 47} and y = {108, 121, 98, 124, 124, 122, 112, 114, 105, 107}

1. Manually compute the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ of a linear model $\hat{y} = \beta_0 + \beta_1 x$ which fits the given data

2. What is the value of MSE calculated between the values of y and the ones returned by the $\hat{y}$ function?

3. Is the trend identified by $\hat{\beta}_1$ significant or it is just due to spurious correlations? You have to provide supporting computations and justifications for your answer.

### 2.6.1 solution

| x | y | $x-\bar{x}$ | $y-\bar{y}$ | $(x-\bar{x})^2$ | $(x-\bar{x})(y-\bar{y})$ |
|---|---|---|---|---|---|
| 45 | 108 | -4.5 | -5.5 | 20.25 | 24.75 |
| 54 | 121 | 4.5 | 7.5 | 20.25 | 33.75 |
| 41 | 98 | -8.5 | -15.5 | 72.25 | 131.75 |
| 55 | 124 | 5.5 | 10.5 | 30.25 | 57.75 |
| 52 | 124 | 2.5 | 10.5 | 6.25 | 26.25 |
| 56 | 122 | 6.5 | 8.5 | 42.25 | 55.25 |
| 49 | 112 | -0.5 | -1.5 | 0.25 | 0.75 |
| 50 | 114 | 0.5 | 0.5 | 0.25 | 0.25 |
| 46 | 105 | -3.5 | -8.5 | 12.25 | 29.75 |
| 47 | 107 | -2.5 | -6.5 | 6.25 | 16.25 |

.

The formula for calculating of the sample mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, so

$\bar{x} = \quad = 49.5$

$\bar{y} = \quad = 113.5$

$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 376.5$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 210.5$

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \quad = 1.7886 \approx$

$\hat{\beta}_o = \hat{y} - \hat{\beta}_1 x == 24.9644 \approx$

## 2.7 ex 2016-02-19

(a) You have a dataset with n = 1000 observations and try to fit different models on the data:

- a linear regression model, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$
- the polynomial regression model, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$
- a smoothing spline (i.e. an even more flexible model than the previous two)

For each of the three models, you calculate both training and test RSS. How would you expect the values of RSS to be (both in the training and in the test case), suppousing that the true relationship between X and Y is (a) linear or (b) cubic?

(b) What is the *additive assumption* in a linear regression model? Show how you would detect a possible interaction between variables and how you would model it. Finally explain, with an example, how the new model would take this interaction into account.

# 3 Classification (8 points)

## 3.1 ex 2015-02-09

Given the two sets of samples from classes RED = {(1, 3), (2, 2), (3.5, 1), (5, 4), (1.5, 4), (4, 2)} and GREEN = {(2, 3), (3, 0.5), (4, 3), (3.5, 2), (1, 2), (2, 1)}, and the three unclassified elements a = (4, 1), b = (2, 3.5), and c = (3, 4)

1. Use the KNN approach to classify the unknown items for k = 1, 3, 5. Apply the Euclidean distance as a metric (note that you can skip the actual distance calculations if you can tell the nearest neighbours at a glance).

2. Describe in details the Linear Discriminant Analysis and Logistic Regression techniques for classification and discuss how these techniques perform, in general, with respect to the KNN classifiers.

## 3.2 ex 2015-02-23

Consider the following dataset with three classes and the Linear Discriminant Analysis model (LDA hereafter)

| X1 | X2 | Class |
|---|---|---|
| 1 | 1 | A |
| 2 | 2 | A |
| 2 | 3 | A |
| 3 | 1 | A |
| 4 | 1 | A |
| 1 | 4 | B |
| 2 | 5 | B |
| 3 | 4 | B |
| 4 | 5 | B |
| 4 | 3 | C |
| 6 | 1 | C |
| 6 | 2 | C |
| 6 | 3 | C |

1. Compute the parameters of a LDA classifier

2. Compute the discriminant functions for the LDA classifier

3. Compute the equations of the boundaries between the three classes according to the classifier

4. Draw the dataset and the boundaries between the classes according to the classifier

5. Compute the parameters in case of a Quadratic Discriminant Analysis model

### 3.3 ex 2015-07-06

(a) Given the subset of Iris dataset in figure, classify the three points identified with white triangles (at coordinates (5.85, 2.3), (4.7, 2.4), and (6.5, 2.5) respectively), using the KNN algorithm with k = 3, 5, 7. Note: if your point has the same amount of neighbors for each class, you can assign it the class of the closest one.

(b) Explain what the curse of dimensionality is. How would you address this problem? {PIC1}

### 3.4 ex 2015-09-14

(a) What is the difference between Discriminative and Generative methods for classification? Provide one example for each category explaining why it can be considered as Discriminative/Generative.

(b) Provide a detailed description of the Generative method you previously introduced, the underlining model, the training algorithm, and derive the shape of its decision boundary.

(c) Provide a detailed description of the Discriminative method you previously introduced, the underlining model, the training algorithm, and describe the shape of its decision boundary.

### 3.5 ex 2015-09-30

(a) Given the dataset in figure, classify the three points identified with white triangles (at coordinates (21, 5), (24, 20), and (26.5, 11.5) respectively), using the KNN algorithm with k = 2, 3, 5. Note: if your point has the same amount of neighbors for each class, you can assign it the class of the closest one.

(b) A hospital collected data for a group of patients to study the relationship between Heart-attack Risk Index (HRI = $X_1$ ), weekly hours of physical activity (PHY = $X_2$), and the probability Y of having a heart attack. Roughly, for a heart attack probability to be low the HRI should be below 5, and the more hours one spends exercising the better it is.

After fitting a logistic regression, the following coefficients were estimated: $\hat{\beta}_0$ = –9.7, $\hat{\beta}_0$ = 1.05, and $\hat{\beta}_2$ = – 0.29.

- estimate the probability for a patient with HRI=5 and PHY=2 to have a heart attack;
- estimate how many hours of PHY a patient with HRI=7.5 should do to have that same probability.

{PIC1}

## 3.6 ex 2016-02-03

Let consider the Linear Discriminant Analysis (LDA) classifier and the following dataset:

$$x = \{3, -1, -4, 0, 2, 5, -1, -2, -2, -2\},$$
$$y = \{A, B, C, A, A, B, A, B, B, C\}$$

- Describe the LDA model and its underlining assumptions. What is its relationship with the Bayes classifier?
- Derive the analytical form of the decision boundary defined by the LDA classifier in a single dimension setting
- Derive the analytical form of the decision boundary defined by the LDA classifier in a multidimensional setting
- Learn an LDA classifier from the provided dataset and compute its classification error on the training set

## 3.7 ex 2016-02-19

In an attempt to convince students to regularly attend his classes, a professor and his TA collected data from their own students and studied the relationship between Class Hours (CH $= X_1$, i.e. the total number of hours attended in class for a given subject), Study Hours (SH $= X_2$, i.e. the total number of hours spent at home studying for that subject), and the probability Y of passing the final exam. Roughy, the more hours a student spends on a subject the higher the probability, but class hours tend to be worth more than the ones spent at home. After fitting a logistic regression, the following coefficients were estimated: $\hat{\beta_0} = -8.75$, $\hat{\beta_1} = 0.25$ and $\hat{\beta_2} = 0.1$.

(a) Estimate the probability for a student with CH = 35 and SH = 20 to pass the exam

(b) Estimate how many hours of SH a student who could attend only CH = 25 hours of classes needs to study to have that same probability to pass the exam

According to Statistical Desicion Theory, the lowest error for a classifier is the Bayes Error, i.e., the error, obtained by the Bayes Classifier, i.e., the classifier which selects the class according to

$$\arg\max_j P(Y = j | X = [x_1, x_2, ..., x_p])$$

(c) Under which posterior distributions does the Logistic Regression classifier obtains the lowest average error rate in the case of (i) binary classes and in the case of (ii) multiple classes (e.g., K classes)?

(d) What is the expected average error for the Bayes Classifier?

## 3.8 ex 2016-07-06

Let's consider 4 "classical" classification algorithms: K-Nearest Neighbours (KNN), Logistic Regression (LR), Linear Discriminant Analysis (LDA), and Suport Vector Machines (SVM)

1. Provide a short description of each of the 4 algorithms highlighting the idea behind each of them, the basics assumptions, the complexity of the decision boundary, the learning process
2. When should we choose each of them?
3. Describe what are the characteristics that could make each of the algorithms preferable to the others.
4. Which of the four algorithms has a non-linear decision boundary? In case of linear decision boundary, how it is possible to extend the algorithm so to have a non-linear one?

# 4 Clustering (8 points)

## 4.1 ex 2015-02-09

Suppose you want to evaluate the results of some clustering algorithms using SSE and Accuracy.

1. Which of these measures is defined as internal, which is external, and what does this mean?

2. After running your function, you obtain the result SSE=21.5. How can you evaluate whether this is a good or bad result? What would you compare this result with?

3. One of the clustering algorithms allows you to choose the number of clusters in advance. You calculate SSE after different executions of this algorithm, using K=2, 3, ..., 10. SSE for K=10 provides the lowest value: what can you deduce from this?

4. Now suppose you have ground truth for your dataset. You run two different clustering algorithms on the same data and obtain the following results:

| | SSE | Accuracy |
|---|---|---|
| Algorithm1 | 115.3 | 87% |
| Algorithm2 | 1285 | 95% |

   What is the meaning of these results? Which algorithm is better?

## 4.2 ex 2015-02-23

Given the two datasets shown in figure (where blue diamonds are data points and red circles different centroids starting points), calculate and show the different steps of the K-Means algorithm for both the examples in the following way:

- At each step, specify the initial positions of the centroids
- Without actually calculating it (unless it is needed to verify distances you cannot tell apart at a glance), for each step specify which centroid the various dataset points belong to
- After you have assigned data points to the different centroids, calculate their new positions and proceed to next step

Tell how many iterations the algorithm needs to converge, compare its behavior in the two cases and write a comment about it: is the final situation the one you might expect/desire? If not, explain why.

{PIC1 } {PIC2 }

## 4.3 ex 2015-07-06

Given the following algorithms:

1. K-means
2. Hierarchical
3. Mixture of Gaussians
4. DBSCAN
5. K-medoids
6. Fuzzy C-means
7. Jarvis-Patrick

complete the following sentences matching them with one (or more!) of the algorithms, answering the questions in parentheses and providing detailed explanations to motivate your choices. (NOTE: although sentences refer to a single algorithm, there may be more than one valid choice. In these cases, provide and motivate all of them).

(a) This algorithm relies on a self-scaling neighborhood (what does this mean? How can this be accomplished?)

(b) This algorithm builds new clusters by merging or splitting existing ones (describe the differences between the two approaches and provide the computational complexity of this approach)

(c) This algorithm provides a soft classification (what does this mean?)

(d) This algorithm can provide good results even if noise is present in the dataset (is it also able to detect which points are noise?)

## 4.4 ex 2015-09-14

Given the dataset shown in figure, execute the steps of a hierarchical (agglomerative) algorithm using the single linkage technique, showing the new links you create at each step of the algorithm (labeling them with a number) and stopping when you obtain two clusters.

{PIC1}

Then, calculate and show the different steps of a K-Means algorithm run on the same dataset with the starting centroid positions $C_1 = (1, 1)$ and $C_2 = (5, 3.5)$, in the following way:

- at each step, specify the initial positions of the centroids

- without actually calculating it (unless it is needed to verify distances you cannot tell apart at a glance), for each step specify which centroid the various dataset points belong to
- after you have assigned points to the different centroids, calculate their new positions and proceed to next step

Are there any differences between the two algorithms? If so, how could you explain their different behavior?

## 4.5 ex 2015-09-30

a) Hierarchical clustering is not a single algorithm but rather a family of different clustering algorithms. Explain (1) how this family is composed, (2) how these algorithms work, and (3) what metrics exist to measure the distance between clusters.

b) Invent a clustering problem (for example, clustering of students according to their grades, news articles according to the words they contain, or images according to their visual descriptors). Describe the problem in detail, specifying e.g. what kind of application you are doing clustering for, the dataset size and dimensionality, what problems you might have while clustering, and so on. Then choose any two of the algorithms we have studied, and try to sell us one of the two, describing the characteristics of both and explaining why using one is better than the other (for instance, in terms of speed, quality of results, etc.).

## 4.6 ex 2016-02-03

From the very definition of what clustering is, we learned that the main purpose of a clustering algorithm is "to group objects in classes, so that inta-class similarity is maximized and inter-class similarity is minimized". Answer the following questions, *providing examples* to support your claims.

- What is the relationship between *similarity* and *distance*? Is it always possible to calculate one from the other?
- Do all clustering algorithms just need similarities/distances to work or are there any other conditions to be met (or parameters to be specified)?
- The quality of a clustering algorithm is often evaluatedin terms of SSE. What is it? What value of SSE is a good value? Is this measure applicable across different datasets (i.e. is SSE = 5 on dataset A always better than SSE = 10 on dataset B)? Explain why.
- Suppose you run the same clustering algorithm on 1000 datasets which were randomly generated within a given interval. The 1000 SSE values you calculate fit a normal distribution with mean $\mu$. You then run the same algorithm on a real dataset (whose data span the same interval you used to generate the random data)

and get a result whose SSE is much smaller than $\mu$. What can you conclude from this?

## 4.7 ex 2016-02-19

Given the two figure below (where blue diamonds are points from the same dataset and red dots ones different centroids starting points), calculate and show the the different steps of the K-Means algorithm for each of the examples in the following way:

- At each step, specify the initial positions of the centroids
- Without actually calculating it (unless it is needed to verify distances you cannot tell apart at a glance), for each step specify which centroid the various dataset points belong to
- After you have assigned points to the different centroids, calculate their new positions and proceed to next step

**NOTE: in both figures, in (1,1), you have a diamond AND a dot!!!**

Tell how many iterations the algorithm needs to converge, compare its behaviour in the two cases and write a comment about it: is the final situation the one you might expect/desire? if not, explain why.

{PIC1} {PIC2}

## 4.8 ex 2016-07-06

K-Means is a clustering algorithm that, despite some limitations, is still widely used for many applicatons.

1. Describe K-Means algorithm in details, and eloborate about its initialization, i.e., what approach would you suggest to address the fact that the result of K-Means clustering depends on the initial positions of centroids?
2. Highlight the main advantages of using K-Means instead of another clustering algorithm (you can explicitly compare K-Means with other algorithms you choose) and suggest some applications for which you consider it better suited.
3. What clustering algorithm would you suggest to address K-Means limit of not being able to deal with non-globular clusters? Choose one (if there are many) and motivate your answer with respect to K-Means.
4. What approach would you suggest to address the need of knowing the number of clusters in advance in K-Means? What alternative clustering algorithm would you choose not to have the issue of selecting an initial number of clusters (discuss your answer).