

POLITECNICO DI MILANO — COMO CAMPUS



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

xesh

Exams 2014-2015

Summary

Project Repository

Click

Team Members

ID	Surname	Name
10460625	Golubeva	Svetlana

Contents

1	Statistical learning (8 points)	3
1.1	ex 2015-02-09	3
1.2	ex 2015-02-23	3
1.3	ex 2015-07-06	4
1.4	ex 2015-09-14	4
1.5	ex 2015-09-30	4
2	Linear regression (8 points)	4
2.1	ex 2015-02-09	4
2.2	ex 2015-02-23	5
2.3	ex 2015-07-06	5
2.4	ex 2015-09-14	5
2.5	ex 2015-09-30	6
3	Classification (8 points)	6
3.1	ex 2015-02-09	6
3.2	ex 2015-02-23	6
3.3	ex 2015-07-06	7
3.4	ex 2015-09-14	7
3.5	ex 2015-09-30	7
4	Clustering (8 points)	8
4.1	ex 2015-02-09	8
4.2	ex 2015-02-23	8
4.3	ex 2015-07-06	9
4.4	ex 2015-09-14	9
4.5	ex 2015-09-30	9

List of Tables

List of Figures

1 Statistical learning (8 points)

1.1 ex 2015-02-09

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.

1.2 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.

1.3 ex 2015-07-06**1.4 ex 2015-09-14**

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

1.5 ex 2015-09-30**2 Linear regression (8 points)****2.1 ex 2015-02-09**

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

2.2 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

- 1.
- 2.
- 3.
- 4.

2.4 ex 2015-09-14

- 1.
- 2.
- 3.
- 4.

2.5 ex 2015-09-30

- 1.
- 2.
- 3.
- 4.

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

- 1.
- 2.
- 3.
- 4.

3.4 ex 2015-09-14

- 1.
- 2.
- 3.
- 4.

3.5 ex 2015-09-30

- 1.
- 2.
- 3.
- 4.

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 $K=2, 3, \dots, 10$. 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

- 1.
- 2.
- 3.
- 4.

4.4 ex 2015-09-14

- 1.
- 2.
- 3.
- 4.

4.5 ex 2015-09-30

- 1.
- 2.
- 3.
- 4.