Homework 6 Basic Machine Learning

For the deadline see Canvas

Version: Thu 11$^{\text{th}}$ Oct, 2018 at 08:04.

Introduction

1. Each group has to submit a **pdf** with their answers and explanation. **Please put your names and group number at the top of the hand in**.

2. For questions about this homework assignment use the Discussion Board on Canvas.

3. Of course you may use a calculator or a programming environment such as Matlab or Python. **But your report should not contain any code. Explain your computations and results in English!**

4. It is allowed to incorporate handwritten notes or derivations or drawings in your submission as long as these are readable!

5. Explain your answers!

Tools and Data files

You may use your Python code to calculate the answers for the assignments, but the report should explain your approach in natural language and if needed in math formulas and should contain your results, of course including explanations! The available data files are in Matlab format. These can be loaded in Python, see scipy tutorial on io. For computing the Principal Components one can use the Singular Value Decomposition (SVD) in Python: scipy tutorial on SVD. The columns of the returned matrix U are the principal components (eigenvectors) and the diagonal values of V the corresponding eigenvalues.

Exercise 1: Constrained Optimization (10 points)

Part a

Solve the minimization problem $f(x_1, x_2) = x_1^2 + 3x_2^2 - 2x_2$ given the constraint $g(x_1, x_2) = x_1 + 3x_2 + 2 = 0$ using Lagrange multipliers.

Part b

Given another constraint $h(x_1, x_2) = -2x_1 + 2x_2 + 1$. Solve the minimization problem for f given that $g(x_1, x_2) \geq 0$ and $h(x_1, x2) \leq 0$. Which constrains are active?

Exercise 2: Probabilistic Classification (10 points)

Consider the data file assignment6_2.mat (can be found on Canvas). This data file contains two matrices A and B which corresponds to 2-dimensional data points for class A and class B, each column represents a data point.

Part a

Estimate the mean and covariance of class A and class B.

Part b

Suppose one has a new data point $\mathbf{x} = (2, 1)$. How will $\mathbf{x}$ be classified if one assumes that A and B are normally distributed?

Part c

Calculate and plot the ROC curve of the classifier on the training set. Assume that class B is the positive class.

Exercise 3: Probabilistic classification and Principal Component Analysis

Consider data file assignment6_3.mat (can be found on Canvas). This data file contains two matrices A and B which corresponds to 3-dimensional data points for class A and class B, each column represents a data point.

Part a

Compute the covariance matrix and the determinant of the estimated covariance matrix for class A. Can one compute $P(x|A)$ if one assumes that class A is normally distributed?

Part b

Compute the Principal Components for the total dataset $X = A \cup B$. Given a new datapoint $\mathbf{x} = (1, 2, 1)$, compute the projection $\mathbf{x}_1$ of $\mathbf{x}$ on this first principal component (principal component with the largest eigenvalue).

Part c

Let A_1 be the projection of A on the first principal component, and Let B_1 be the projection of B on the first principal component. Compute the mean and covariance for both A_1 and B_1.

Part d

Let $\mathbf{x}_1$ be the projection of $\mathbf{x}$ on the first principal component. Compute $P(\mathbf{x}_1|A_1)$ and $P(\mathbf{x}_1|B_1)$, assuming that both A_1 and B_1 are normally distributed. How will $\mathbf{x}$ be classified if only looks at the first principal component?

Exercise 4: Breast cancer case. (20 points)

Download the breast cancer data set breast_cancer.zip (available on Canvas). This zip-file consists of a training, set, validation set and an explanation (breast-cancer-description.txt). You are asked to implement several classification algorithms using a probabilistic Bayesian approach using (multi-variate) Gaussians probability distributions. To summarize the description of the data set, there are 10 inputs and 2 classes:

1. Sample code number
2. Clump Thickness
3. Uniformity of Cell Size
4. Uniformity of Cell Shape
5. Marginal Adhesion
6. Single Epithelial Cell Size
7. Bare Nuclei
8. Bare Nuclei
9. Normal Nucleoli
10. Mitoses
11. Class (2 for benign, 4 for malignant)

The features 2 up to and including 10 all have integer values between 1 and 10.
The data set is split in a training set (60%) and a validation set (40%), this split is stratified (priors for the train and validation set are equal).

Part a

In order investigate the effect of dimensionality reduction and PCA apply the Bayesian approach for each of the following set of features:

1. x4 (Uniformity of cell shape) only
2. First principal component
3. x4 and x8 (Mitoses) only
4. First two principal components
5. The principal components which explain 80% of the variance
6. All inputs (except of course the sample code number)

Thereby implementing univariate, bivariate, and multi-variate versions of the classifier. Remember that for training or estimating PCAs you may only use the training set! For each choice:

- Estimate and report relevant parameters for both classes.
- Report the errors on the training and validation set, give the confusion matrix , and precision and recall for the cancer class malignant.

Exercise 5

Part a

Apply logistic classification to the breast cancer data set above. Report the errors on the training and validation set, give the confusion matrix , and precision and recall for the cancer class malignant.
Do not forget to add a 1 to the input and to adapt the output labels to 0 and 1!.

Part b

Calculate and plot the ROC curve of your logistic classifier on the validation set. Assume that “malignant” is the positive class.

Part c

Compare with the approaches in Exercise 3. Which classification model is the best and why.