TDA 231 Machine Learning: Homework 0

Goal: Introduction to Probability, Matlab Primer
Grader: Mikael Kågebäck

Due Date: January 25, 2016

General guidelines:

1. All solutions to theoretical problems, and discussion regarding practical problems, should be submitted in a single file named *report.pdf*
2. All matlab files have to be submitted as a single zip file named *code.zip*.
3. The report should clearly indicate your name, personal number and email address
4. All datasets can be downloaded from the course website.
5. All plots, tables and additional information should be included in *report.pdf*

1 Theoretical problems

Problem 1.1 [Bayes Rule, 5 points]

After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a very serious cancer and that the test is 99% accurate i.e. the probability of testing positive given you have the disease is 0.99. The probability of testing negative if you don't have the disease is the same. The good news is that it is a very rare condition affecting only 1 in 10,000 people. What is the probability you actually have the disease? (Show all calculations and the final result.)

Problem 1.2 [Correlation and Independence, 5 points]

Let X be a continuous variable, uniformly distributed in $[-1, +1]$ and let $Y := X^2$. Clearly Y is not independent of X – in fact it is uniquely determined by X. However, show that $\text{cov}(X, Y) = 0$.

2 Practical problems

Useful matlab functions:

- *General:* arrayfun, (anonymous functions using @), min, max, mean, cov, inv
- *Plotting:* plot, scatter, ezplot, legend, hold, grid title, saveas

Problem 2.1 [Plotting normal distributed points, 5 points]

Generate 1000 points from 2D multivariate normal distribution having mean $\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and covariance $\Sigma = \begin{bmatrix} 0.1 & -0.05 \\ -0.05 & 0.2 \end{bmatrix}$. Define the function $f(\mathbf{x}, r) := \frac{(\mathbf{x}-\mu)^\top * \Sigma^{-1} * (\mathbf{x}-\mu)}{2} - r$. On a single plot, show the following:

- The level sets $f(\mathbf{x}, r) = 0$ for $r = 1, 2, 3$.
- Scatter plot of randomly generated points with points lying outside $f(\mathbf{x}, 3) = 0$ showing in black while points inside shown in blue.
- Title of the plot showing how many points lie outside $f(\mathbf{x}, 3) = 0$.

Submit your final plot as well as your implementation.

Problem 2.2 [Covariance and correlation, 5 points]

Load dataset0.txt (X) containing 1074 data points each having 12 features related to US schools. Compute the covariance and correlation matrix for X. Scale each feature in X between $[0, 1]$ to obtain a new dataset Y. Compute the covariance and correlation matrices for X and Y, and plot them (e.g. as colormaps). What do you observe? Show a scatter plot of the pair of features in Y having minimum correlation, indicating in the title the feature indices and the correlation value. Submit the plot and your implementation.