582631 Introduction to Machine Learning, Autumn 2013
Exercise set 4 (based on third week of lectures)

To be discussed in the exercise session 22 November.

Credit is given based on the written solutions turned in to the course assistant. Extra credit (0.5 points per problem) is given for willingness to present the solution to the pen-and-paper problems at the exercise session.

The deadline for handing in your solutions is **9:00am on Wednesday, 20 November.**

Send your solutions to the course assistant (Yuan.Zou@cs.helsinki.fi) by e-mail. You should send one PDF file including your solutions to the pen-and-paper problems and the report for the programming problem, and a single compressed file including the code for the programming problem.

Pen-and-paper problems

Problem 1 (6 points)

Assume that in a classification problem the classes have marginal distribution $P(Y)$ given by $P(Y = 0) = 0.4$, $P(Y = 1) = 0.3$, and $P(Y = 2) = 0.3$, and class-conditional distributions $P(\mathbf{X} \mid Y)$ as specified below (with $\mathbf{X} = (X_1, X_2)$):

		X_1 0	X_1 1
	0	0.2	0.1
X_2	1	0.4	0.2
	2	0.0	0.1

$P(X_1, X_2 \mid Y = 0)$

		X_1 0	X_1 1
	0	0.6	0.1
X_2	1	0.1	0.1
	2	0.1	0.0

$P(X_1, X_2 \mid Y = 1)$

		X_1 0	X_1 1
	0	0.1	0.4
X_2	1	0.3	0.0
	2	0.2	0.0

$P(X_1, X_2 \mid Y = 2)$

In our classification problem the prediction type is 'forced choice' and the cost function that we seek to minimize is classification error (i.e. unit cost for any wrong answer, zero cost for correct answers).

(a) Give the optimal Bayes classifier, i.e. for each of the 6 possible values of $\mathbf{X}$, what is the best forced-choice prediction of Y?

(b) Compute the Bayes Error, i.e. the error rate (percentage of wrong answers) for the optimal Bayes classifier.

(c) Construct the corresponding Naïve Bayes classifier, i.e. when approximating the class-conditional distributions by the product of the marginals, what is the best forced-choice prediction of Y for each of the 6 possible values of $\mathbf{X}$?

(d) Compute the error rate of the Naïve Bayes classifier.

(e) Is the error rate of the Naïve Bayes classifier higher or lower than the Bayes Error? Could you have answered this last question without computing the actual values?

Problem 2 (3 points)

Consider a binary classification problem with two classes, $y = 0$ and $y = 1$, with equal prior probability, i.e. $P(y = 0) = P(y = 1) = 0.5$, and class-conditional distributions

$$\begin{aligned} p(\mathbf{x} \mid y = 0) &= \mathcal{N}(x_1;\ 0, \sigma_0^2)\mathcal{N}(x_2;\ 0, \sigma_0^2) \\ p(\mathbf{x} \mid y = 1) &= \mathcal{N}(x_1;\ 0, \sigma_1^2)\mathcal{N}(x_2;\ 0, \sigma_1^2), \end{aligned}$$

where $\mathbf{x} = (x_1, x_2)^T$ contains the two continuous-valued attributes and the variances are set to $\sigma_0^2 = 1$ and $\sigma_1^2 = 16$. Calculate the decision boundary for the optimal (Bayesian) classifier which minimizes the expected classification error for new, unseen data. Compute a numerical estimate of the Bayes error, i.e. the expected error of the optimal classifier, by computer simulation, as follows: Generate 10 000 samples from the above model (for each sample, first randomly select $y = 0$ or $y = 1$ with equal probability, then draw $\mathbf{x}$ from the appropriate $p(\mathbf{x} \mid y)$), classify with your analytically computed optimal Bayes classifier, and compute the resulting error rate. (If you wish to, you can also try to analytically compute the Bayes error, but this is not required to get the full points as it may require some math or statistics knowledge over and above the course prerequisites.)

Computer problem

General instructions:

- Return a brief written report (as PDF, included in the same file as your pen-and-paper solutions) and one directory (as a zip or tar.gz file) including all your code.
- Do not include any of the data files in your solution file.
- Always mention your name and student ID in the report.
- We use the report as the main basis for grading: All your results should be in the report. We also look at the code, but we won't however go fishing for results in your code.
- The code needs to be submitted as a runnable file or set of files (command to run it given in the report).
- In your report, the results will be mostly either in the form of a figure or program output. In both cases, add some sentences which explain what you are showing and why the results are the answer to the question.
- If we ask you to test whether an algorithm works, always give a few examples showing that the algorithm indeed works

Problem 3 (15 points)

In this problem we will attempt to reproduce Figure 2.4 in Hastie et al (2009)[1], also presented in the slides of lecture 6, showing the error on the test and training data as a function of the parameter k in a kNN classifier, for the binary classification problem defined in Problem 2 (above).

(a) First, randomly draw 500 datapoints $(\mathbf{x}, y)$ from the distribution specified in Problem 2 above. This will be your *training* set. Plot these points in a two-dimensional scatterplot with red points indicating class $y = 0$ and blue points indicating class $y = 1$. You should see a spherically symmetric pattern where the red points cluster around the origin while the blue points are more spread out. (You probably want to plot the blue points below the red ones, to make the pattern clear.)

(b) Next, randomly draw 2000 new datapoints from the same distribution. This will be your *test* set. Plot these points in a separate plot using the same technique.

(c) For each $k \in \{1, 3, 5, 7, 9, 13, 17, 21, 25, 33, 41, 49, 57\}$, apply a kNN classifier to classify the points in the test set. That is, for each point in the test set, find its k closest (by Euclidean distance) neighbors from the training set, and estimate the class of the test set example by the mode of the classes of those neighbors. For each value of k, plot the percentage of misclassifications (number of errors divided by the size of the test set), similarly to Figure 2.4 in Hastie et al (2009).

(d) For the same values of k as in (c), apply the kNN classifier to classify the points in the training set. That is, for each point in the training set, find its k closest (by Euclidean distance) neighbors from the training set, and estimate the class of the training set example by the mode of the classes of those neighbors. Similarly to in (c), plot the percentage of misclassifications, into the same plot as in (c) for easy comparison.

(e) Draw the estimated Bayes error (from Problem 2 above) as a line into the same figure.

Hints: The plot as a whole should look qualitatively similar to that of Hastie et al: The test error of kNN is minimized for intermediate values of k, while the training error is optimized (and is equal to zero in fact!) for $k = 1$. The minimum of the test error should be relatively close to the Bayes error.

[1] `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`