COMP4670/6467 - 2013 Semester 1

Introduction to Statistical Machine Learning

Assignment 1

Maximum marks	20
Weight	20% of final grade
Submission deadline	Monday, April 22, 2013, 23:59
Document format	Portable Document Format (PDF); ASCII file for Pyton code. Please package multiple files into a .zip or .tar archive. Put your name and students ID on the top of each document.
Submission mode	e-mail to christfried.webers@nicta.com.au
Formula Explanations	All formulas which you derive need to be explained unless you use very common mathematical facts. Picture yourself as explaining your arguments to somebody who is just learning about your assignment. With other words, do not assume that the person marking your assignment knows all the background and therefore you can just write down the formulas without any explanation. It is your task to convince the reader that you know what you are doing when you derive an argument.
Code quality	Python code should be well structured, use meaningful identifiers for variables and subroutines, and provide sufficient comments. Please refer to the examples given in the tutorials.
Code efficiency	An efficient implementation of an algorithm uses fast subroutines provided by the language or additional libraries. For the purpose of implementing Machine Learning algorithms in this course, that means using the appropriate data structures provided by Python and in numpy/scipy (e.g. Linear Algebra and random generators).
Late Penalty	20% per day overdue (a day starts at midnight!)
Cooperation	All assignments must be done individually. Cheating and plagiarism will be dealt with in accordance with University procedures (please see the ANU policies on "Academic Honesty and Plagiarism" `http://academichonesty.anu.edu.au`). Hence, for example, code for programming assignments must not be developed in groups, nor should code be shared. You are encouraged to broadly discuss ideas, approaches and techniques with a few other students, but not at a level of detail where specific solutions or implementation issues are described by anyone. If you choose to consult with other students, you will include the names of your discussion partners for each solution. If you have any questions on this, please ask the lecturer before you act.
Solutions	To be presented in the tutorials.

1 Probabilities

1.1 (1/20) Covariance of Sum

Prove that the following holds for the variance of a sum of two random variables X and Y

$$\text{var}[X+Y] = \text{var}[X] + \text{var}[Y] + 2\,\text{cov}[X,Y],$$

where cov[X,Y] is the covariance between X and Y.
For each step in your proof, provide a verbal explanation why this transformation step holds.

1.2 (2/20) Probability of Babies

My neighbour has two children. Let us assume that the gender of a child is a binary variable (in reality it is not!) following a Bernoulli distribution with parameter 1/2 and that those defined genders of children are iid.

1. How many girls/boys will my neighbour most likely have?
2. Suppose I ask him whether he has any girls, and he says yes. What is the probability that one child is a boy?
3. Suppose instead that I happen to see one of his children in the garden, and it is a girl. What is the probability that the other child is a boy?

1.3 (3/20) Maximum Likelihood for Multivariate Gaussian Distribution

Assume input data $\mathbf{x_n} \in \mathbb{R}^D$, $n = 1, \ldots, N$, are drawn i.i.d. from a Gaussian distribution.

1. Define the likelihood of drawing all data $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N})^T$ given the parameters of the Gaussian distribution.
2. Calculate the parameters of the Gaussian distribution for which the above defined likelihood has an extremum.
3. Show that with respect to the mean, the extremum found is a maximum.
4. Is the order in which the maximising parameters are found arbitrary? Please justify your answer.
5. How do the parameters maximising the likelihood change if the order of the input data $\{\mathbf{x_n}\}$, $n = 1, \ldots, N$ changes? Justify your result.

Hints: The directional derivative of the determinant $|\mathbf{A}|$ of a matrix $\mathbf{A}$ in direction $\mathbf{B}$ is given by

$$\mathcal{D}|\mathbf{A}|(\mathbf{B}) = |\mathbf{A}|\,\text{tr}\left\{\mathbf{A}^{-1}\mathbf{B}\right\}$$

where $\text{tr}\{\cdot\}$ is the trace function.
Furthermore, the directional derivative of the inverse $\mathbf{A}^{-1}$ of a matrix $\mathbf{A}$ in direction $\mathbf{B}$ is given by

$$\mathcal{D}\mathbf{A}^{-1}(\mathbf{B}) = -\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}.$$

1.4 (2/20) Lifetime of Equipment

The lifetime X of some equipment is modelled by an exponential distribution with an unknown parameter θ as

$$p(x \mid \theta) = \theta e^{-\theta x} \qquad x \geq 0, \qquad \theta > 0.$$

1. Calculate the maximum likelihood estimator (MLE) $\widehat{\theta}$ for a number of observed iid lifetimes $X_1, \ldots, X_N$.
2. Derive the relation between the mean of $X_1, \ldots, X_N$ and the maximum likelihood estimator $\widehat{\theta}$.
3. Suppose we observe $X_1 = 5$, $X_2 = 4$, $X_3 = 3$ and $X_4 = 4$ as the lifetimes (in years) of four different machines. What is the MLE given this data?

2 Decision Theory

2.1 (2/20) Lower Bound for the Correct Classification

Consider two nonnegative numbers a and b. Show that, if $a \leq b$, then $a \leq (ab)^{1/2}$ holds.

Now, consider a two-class classification problem where the decision region was chosen to mimimise the probability of misclassification. Use the above inequality to prove that

$$p(\text{mistake}) \leq \int \sqrt{p(\mathbf{x}, \mathcal{C}_1)}\, \sqrt{p(\mathbf{x}, \mathcal{C}_2)}\, \mathrm{d}\mathbf{x}.$$

Please explain each step in your derivation. If possible, provide the intuition behind a step.

3 Dimensionality Reduction

3.1 (5/20) Projection with Fisher's Discriminant

Fisher's Linear Discriminant finds a projection of the input data in which two goals are combined: maximising the distance between the centres of points belonging to each class and minimising the variance of data in each class.

We will investigate Fisher's Linear Discriminant using the Iris Flower data data set in order to project the data from the original 4 dimensions into a lower dimension $D' < 4$.

1. Given the set of input data $\mathbf{X}$ and class labels $\mathbf{t}$ of K different classes, calculate the within-class and between-class scatter matrices $\mathbf{S}_W$ and $\mathbf{S}_B$ (see lecture slides). (Note that scatter matrices are not estimates of covariance matrices because they differ by a scaling factor related to the number of data points.) Find the matrix $\mathbf{W}$ with columns equal to the D' normalised eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$ which are associated with the D' largest eigenvalues.

2. For $D' = 2$:

   (a) Report the two eigenvalues and eigenvectors found.

   (b) Provide a plot of the projected data using different colours for each class.

   (c) Discuss the ratio of the two eigenvalues found with respect to the task of classifying the data in the projected 2-dimensional space.

3. For a set of N projected data $\mathbf{Y} \in \mathbb{R}^{N \times D'}$, implement code to calculate the criterion J given by

$$J = \mathrm{tr}\left\{\mathbf{s}_W^{-1}\,\mathbf{s}_B\right\}$$

   using the definitions

$$\mathbf{s}_W = \sum_{k=1}^{K} \sum_{n \in \mathcal{C}_k} (\mathbf{y}_n - \boldsymbol{\mu}_n)(\mathbf{y}_n - \boldsymbol{\mu}_n)^T \qquad \mathbf{s}_B = \sum_{k=1}^{K} N_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{y}_n \qquad \boldsymbol{\mu} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{y}_n$$

   where $\mathcal{C}_k$ is the set of indices of data belonging to class k, and N_k is the number of data points belonging to class k.

4. Project the Iris data into $D' = 2$ dimensions using the $\mathbf{W}$ found in 2. and report the criterion J for this projection. Using the original data of the Iris Flower data set, report the criterion for all 2-dimensional orthogonal projections onto a plane spanned by a pair of axes out of the 4 axes of the data set. Compare all criteria J found and discuss the results.

4 Cross Validation and Classification

In the next problems, we will use the Iris flower data set available via the course web site at

`http://sml.nicta.com.au/isml13/assignments/bezdekIris.data.txt`

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by Sir Ronald Aylmer Fisher (1936) as an example of discriminant analysis. (The file bezdekIris.data.txt corrects two erroneous entries in the original data set.)

The file consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor).

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	Iris-setosa
...	...	...	...	...
7.0	3.2	4.7	1.4	Iris-versicolor
...	...	...	...	...
6.3	3.3	6.0	2.5	Iris-virginica
...	...	...	...	...

The first four comma separated entries in this data set are the input data $\mathbf{x} \in \mathbb{R}^4$ as floating point numbers, the fifth entry is a class name from the ordered set {"Iris-setosa", "Iris-versicolor", "Iris-virginica"} which you should map to {0, 1, 2}.

4.1 (5/20) k- Nearest Neighbours Algorithm – k-NN

The k-Nearest Neighbours Algorithm (k-NN) classifies a new point in the input space by the most frequent class amongst its k nearest neighbours provided as 'training' data. If more than one class is the most frequent, the class is decided randomly. The neighbourhood of a point in input space is given by a metric, usually the Euclidian metric.

1. Implement the k-NN algorithm using a Euclidian distance for the Iris Flower data set.
2. Apply 2-fold, 5-fold and 10-fold Cross Validation to select the best k for the k-NN algorithm.

   In case of several k having the same lowest error rate, we pick the largest one. Explain why this is a good strategy.
3. Report the results for $k = 2, 4, \ldots, 38, 40$.
4. The optimal errors decrease with the fold number. Explain why.
5. The optimal k increases with the fold number. Explain why.
6. Provide the listing of your program together with the solution.