COMP4670/6467 - 2014 Semester 1

Introduction to Statistical Machine Learning

Assignment 1

Maximum marks	20
Weight	20% of final grade
Submission deadline	Monday, April 21, 2014, 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 (2/20) Medical Test

Assume the prevalence of a certain disease in the general population is 1%. Assume further, there exists a test for the disease, which detects the disease in 80% of the patients having the disease. However, the test also diagnoses a healthy patient as having the disease in 9.5% of the cases.

1. Calculate the probability of a person (randomly selected from the population) really having the disease if the test is positive?

2. Explain your result.

3. Some researchers are aiming to improve the test by reducing the percentage of healthy persons diagnosed with the disease (false positives). Their aim is to guarantee a 50% chance that a person (randomly selected from the population) and tested as positive with the improved test really has the disease. What would the rate of false positives have to become in order to achieve this?

4. In order to make a decision whether a patient has the disease or not, researchers decide to weight the joint probabilities with the following loss matrix **L**

	detected disease	detected healthy
has disease	0	1000
is healthy	1	0

   where the rows correspond to the true classes and the columns correspond to what was detected. Explain possible reasons for choosing the four entries in **L**.

5. Calculate the expected loss for the given problem.

6. Assume now that the decision rule used will minimise the expected loss using the above given loss matrix **L**. Following this rule, which decision (either disease or healthy) will be made, if the result of the test is

   (a) disease, or

   (b) healthy?

   Discuss your result.

1.2 (2/20) Maximum Likelihood (ML) and Maximum A Posteriori (MAP)

a) We assume data samples $X_n = \{x_1, \ldots, x_n\}$ were generated i.i.d. from a uniform distribution $\mathcal{U}(x \mid 0, \theta)$ between 0 and an unknown positive parameter θ:

$$p(x \mid \theta) = \mathcal{U}(x \mid 0, \theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

Assume the data samples $X_4 = \{5, 7, 3, 8\}$ have been observed. Calculate $\theta_{ML} = \arg\max_\theta p(X_4 \mid \theta)$, the maximum likelihood estimate of θ for the observed data.

b) Assume data samples X_n are received one-by-one. Derive the posterior probability $p(\theta \mid X_n)$ from the probability $p(\theta \mid X_{n-1})$.

c) Calculate $p(\theta \mid X_4)$, the posterior distribution of θ given that the data $X_4 = \{5, 7, 3, 8\}$ have been observed and the initial prior is assumed to be $p(\theta) = p(\theta \mid X_0) = \mathcal{U}(x \mid 0, 10)$.

d) Calculate $\theta_{MAP} = \arg\max_\theta p(\theta \mid X_4)$, the *maximum a posteriori* estimate of θ given the data X_4 and the initial prior $p(\theta)$ as in the previous question.

1.3 (3/20) Laplace Approximation

The function

$$f(z) = z^k e^{-z^2/2} \qquad z \in [0, \infty), \qquad k > 0$$

can be considered as an (unnormalised) probability density.

1. Verify that it is possible to approximate $f(z)$ with the Laplace Approximation.
2. Using the Laplace Approximation, find the mean and the variance of the Normal Distribution which best approximates the normalised version of $f(z)$.
3. The analytical form of the normalisation of $f(x)$ is not so easy to find. Therefore, use Python to implement a numerical approximation using $N = 100$ identically sized intervals between 0 and $a = 10$ to calculate the normalisation

   $$\int_0^\infty f(z)\,\mathrm{d}z \approx \int_0^a z^k e^{-z^2/2}\,\mathrm{d}z \approx \sum_{i=1}^{100} \dots$$

   and report the results for the normalisation with a precision of 5 digits after the comma for the three cases $k = \{0.5, 3, 5\}$.
4. Why is it reasonable to replace the upper limit of ∞ with $a = 10$?
5. For each of the three cases $k = \{0.5, 3, 5\}$, plot the normalised function $f(z)$ and the corresponding Normal Distribution with parameters resulting from the Laplace Approximation.

1.4 (3/20) Bayesian Linear Regression

1. Given a nonsingular matrix $\mathbf{A}$ and a vector $\mathbf{v}$ of comparable dimension, prove the following identity:

   $$(\mathbf{A} + \mathbf{v}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} - \frac{(\mathbf{A}^{-1}\mathbf{v})(\mathbf{v}^T\mathbf{A}^{-1})}{1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{v}}. \tag{1}$$

2. In the lecture about Bayesian Linear Regression, the predictive distribution with a simplified prior $p(\mathbf{x} \,|\, \alpha) = \mathcal{N}(\mathbf{x} \,|\, 0, \alpha^{-1}\mathbf{I})$ was given as a Gaussian distribution, $p(t \,|\, \mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t \,|\, \mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))$ with variance

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}).$$

   where $\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}$, and β is the variance of the homogenous Gaussian noise assumed for the data.

   After using another training pair $(\mathbf{x}_{N+1}, t_{N+1})$ to adapt the model, the variance of the predictive distribution becomes

$$\sigma_{N+1}^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_{N+1} \boldsymbol{\phi}(\mathbf{x}).$$

   Show that the variances $\sigma_N^2(\mathbf{x})$ and $\sigma_{N+1}^2(\mathbf{x})$ associated with the predictive distributions satisfy

$$\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x}). \tag{2}$$

   Hint: Use the equality (1) to prove the result.

3. What can you deduce from the inequality (2) ?

2 Dimensionality Reduction

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

```
http://sml.nicta.com.au/isml14/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}, respectively.

2.1 (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 all points belonging to the same 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

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

   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.

3 Cross Validation and Classification

3.1 (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 in the training data. If more than one class is the most frequent, the class is decided randomly.

1. Implement the k-NN algorithm as a function of k, a set of training data and a set of test data, and reporting as error the number of misclassifications of the test data. Use the Euclidian metric for the input space.

2. Implement S-fold cross validation using the k-NN algorithm applied to the Iris Flower data set. For arbitrary choices of S and k, your implementation returns the cross validation error.

3. Run the cross validation for $S \in \{2, 5, 10\}$ and $k \in \{1, 3, \ldots, 37, 39\}$ and report the cross validation error.

   For each S, report the k for which the error is minimal.

   In case of several k having the same lowest error, pick the largest of those k. Explain why this is a good strategy.

4. For each k, how do the errors change with the fold number? Provide an explanation for your result.

5. How does the optimal k change with the fold number? Provide an explanation for your result.

6. Provide the listing of your Python code as a separate file.