COMP4670/6467 - 2013, Semester 1

Introduction to Statistical Machine Learning

Assignment 2

Maximum marks	20
Weight	20% of final grade
Submission deadline	Monday, May 20, 2013, 23:59
Document format	Portable Document Format (PDF); ASCII file for Pyton code
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 about this, please ask the lecturer before you act.
Solutions	To be presented in the tutorials.

1 (1/20) Sampling from a Multivariate Normal Distribution

Many random number generators provide only scalar normal distributed random samples with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}$

$$x \sim \mathcal{N}(x \,|\, \mu, \sigma^2).$$

Using only such a scalar random generator, explain how you can create vectorial data $\mathbf{x} \in \mathbb{R}^n$ distributed as a Multivariate Normal Distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$, i.e.

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, \mathbf{C}).$$

2 (3/20) Reverse Prediction

Given a set of n training data $\mathbf{x}_i \in \mathbb{R}^{n \times d}, i = 1, \dots, n$, and scalar targets $t_i \in \mathbb{R}^n, i = 1, \dots, n$, we denote the optimal $\mathbf{w}$ minimising the quadratic regression error of all data for a linear model

$$\widetilde{t}_i = \mathbf{w}^T \mathbf{x}_i \qquad\qquad i = 1, \dots, n$$

as $\mathbf{w}^\star$.

a) Write down the complete optimisation problem for $\mathbf{w}^\star$.

b) Find a solution for $\mathbf{w}^\star$. Under which conditions is $\mathbf{w}^\star$ unique?

Now consider the problem of mapping targets t_i back to values in the input space, i.e.

$$\widetilde{\mathbf{x}}_i = t_i \mathbf{u}. \tag{1}$$

c) Write down the optimisation problem to minimise the squared error between the training inputs $\mathbf{x}_i$ and the predictions of the model (1), $\widetilde{\mathbf{x}}_i$, for the given targets.

d) Find a solution for $\mathbf{u}^\star$ which minimises the error in c). Under which conditions is $\mathbf{u}^\star$ unique?

e) In case $\mathbf{w}^\star$ and $\mathbf{u}^\star$ happen to be unique : Derive a relation between $\mathbf{w}^\star$ and $\mathbf{u}^\star$.

3 (2/20) Conditional Probability and Variance via Parzen Estimator

Given N data points $x_n \in \mathbb{R}$ and targets $t_n \in \mathbb{R}$, $n = 1, \dots, N$, consider a new input x and target t.

Using a *Parzen density estimator*, the joint probability for the new input and target, $p(x, t)$, can be estimated as

$$p(x,t) = \frac{1}{N} \sum_{n=1}^{N} f(x - x_n, t - t_n)$$

where $f(x, t)$ is called a *component density function.* Assume in the following, that the component density function is an isotropic Gaussian with mean $(0, 0)^T$ and covariance $\sigma^2 \mathbf{I}$ where $\mathbf{I}$ is the two-dimensional identity matrix.

1. Write down the conditional density $p(t \,|\, x)$, the conditional mean $\mathbb{E}\left[t \,|\, x\right]$, and the conditional variance $\operatorname{var}[t \,|\, x] = \int (t - \mathbb{E}\left[t \,|\, x\right])^2 \, p(t \,|\, x) \, \mathrm{d}t$, with the help of a function $k(x, x_n)$.
2. Show that for the function $k(x, x_n)$ the following holds

$$\sum_{n=1}^{N} k(x, x_n) = 1.$$

4 (2/20) Maximum Margin Hyperplane

Given two classes and a data set of only two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^D$, one from each class.

(a) Show that, irrespective of the dimensionality D of the data space, these two points are sufficient to determine the maximum margin hyperplane.

(b) What is the dimension of this maximum margin hyperplane?

(c) Given a new data point $\mathbf{x}$, provide a discriminant function $f(\mathbf{x})$ which classifies this new data point.

(d) Discuss the values of the Lagrange parameters.

5 (2/20) Kernels and Feature Maps

Given data in a 2-dimensional input space, we define a kernel function $k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + 1)^3$ for each pair of input data $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$.

(a) Prove that $k(\mathbf{x}, \mathbf{z})$ is a kernel by explicitly calculating the feature mapping $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x}))^T \in \mathbb{R}^m$ with the smallest dimension m which expresses the function $k(\mathbf{x}, \mathbf{z})$ as an inner product in a m-dimensional feature space: $k(\mathbf{x}, \mathbf{z}) = \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{z})$.

(b) Consider a generalisation in the form of a kernel $k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + 1)^p$ for each pair of input data $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$. Can your method of finding an inner product in a higher dimensional space of dimension m which you developed in (a) also be applied to input space dimensions $n > 2$? If the answer is 'yes', please explain how.

(c) Can the method of finding an inner product in a higher dimensional space of dimension m also be applied to higher powers $p > 3$? If the answer is 'yes', please explain how.

6 (5/20) EM for Trees

A laser-range finding system is used to map the location of trees in a forest. After collecting data, the system provides a set of N measurements $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^2$ of potential tree locations in a 2D cartesian coordinate frame. Assume there are K trees with true centers $\boldsymbol{\mu}_k \in \mathbb{R}^2$, $k = 1, \ldots, K$. Furthermore, assume that the probability that a tree with center $\boldsymbol{\mu}_k$ generated a particular measurement $\mathbf{x}_n$ is a normal distribution $\mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_k, \sigma^2 \mathbf{I})$ with standard deviation σ. (We assume that the trees have a small diameter and thus the offset of measurement from the true center is negligible.)

Answer the following four questions:

(a) Define latent variables for this problem and a joint generative model of the observations and latent variables given the parameters.

(b) Derive the expected log likelihood of this joint model.

(c) Derive the E-step update and show how to compute any required expectations.

(d) Derive the M-step for EM.

7 (5/20) (Semi-/Un-)Supervised Learning and EM

In this section we'll try various density estimation techniques on the Fisher Iris data set (from the course website). Remember, for unsupervised learning you are not allowed to use the class labels (5th column)!

(a) **Supervised Learning:** Fit three 4D Gaussians to the given data (i.e., one 4D Gaussian per class). Show the means and covariance matrices for each class.

Use 10-fold cross-validation to evaluate the classification error of this approach.

(b) **Unsupervised Learning:** Fit a mixture of 3 multivariate Gaussians to the data, using the EM algorithm to fit the means and covariance matrices. Provide a listing of your code.

Produce a table of both the expected log-likelihood and the log-likelihood of the data versus the EM iteration number for 5 restarts from different random initial conditions. Do you find the same solution each time?

(c) **Semisupervised Learning:** If 95% of the data in the Fisher data set was unlabeled, you could use a supervised classifier from part (a) trained on just the labeled data. Can you think of a way to incorporate the unlabeled data in using the EM algorithm (b) to produce a semi-supervised classifier? Describe such an algorithm. Which do you think would perform better? No implementation needed, but defend your answer.