University of Toronto at Scarborough
Department of Computer and Mathematical Sciences

CSCC11: Machine Learning and Data Mining

Midterm exam
Fall 2015

Duration: 50 minutes
No aids allowed

There are 9 pages total (including this page). You may use the back of the pages for rough work, but **only** what is written on the front of the pages will be marked.

Family name: _____

Given names: _____

Student number: _____

Signature: _____

By signing above, I certify that the work contained within is my work and my work alone and understand that copying another students work or allowing another student to copy my work is a serious academic offence.

Question	Marks	
1	_____	/ 15 marks
2	_____	/ 12 marks
3	_____	/ 14 marks
4	_____	/ 12 marks
5	_____	/ 7 marks
Total	_____	/ 60 marks

1. **Short Answer Questions [15 marks]**

   (a) **[3 marks]** What is *Bayes' Rule*? Define it and then use it to write the probability of the parameters θ given the data $\mathcal{D}$. Specify which elements correspond to the *posterior*, the *likelihood*, the *prior* and the *evidence*.

   (b) **[6 marks]** There are multiple ways to estimate parameters θ given data, $\mathcal{D}$. Name and define three such estimators mathematically (that is, write equations).

(c) **[2 marks]** What is the *naïve Bayes'* assumption that is sometimes used in classification? Be as formal and specific as possible.

(d) **[4 marks]** Naïve Bayes' can also be used when estimating the parameters of a Gaussian distribution. In this case, explain what the primary difference is with and without Naïve Bayes'. Be as formal and specific as possible. What parameters are (or are not) estimated in each case?

2. **Gaussian Distributions [12 marks]**

(a) **[1 mark]** A Gaussian random variable $x \in \mathbb{R}$ has *probability density function*

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where μ and σ are parameters of the distribution. Draw this PDF and clearly label how μ and σ impact the the curve.

(b) **[3 marks]** Let x and y be two independent Gaussian random variables with means μ_x and μ_y, variances σ_x^2 and σ_y^2. If z is another random variable with $z = x + y$ then one can show that z also has a Gaussian distribution. Derive its mean $E[z]$ and variance $E[(z - E[z])^2]$.

(c) **[1 marks]** We can also derive the effects of scaling in a similar way. That is if $z = cx$ where c is a known constant, derive the mean and variance of z.

(d) **[3 marks]** Consider two noisy observations x_1 and x_2 of an unknown quantity w

$$\begin{aligned} x_1 &= c_1 w + \eta_1 \\ x_2 &= c_2 w + \eta_2 \end{aligned}$$

where $\eta_1 \sim N(0, \sigma_1^2)$, $\eta_2 \sim N(0, \sigma_2^2)$ and c_1, c_2 are non-zero constants. Further, assume that you have a prior which states that $w \sim N(0, 1)$. Derive the negative log posterior distribution of w given the observations x_1 and x_2 and the constants $c_1, c_2, \sigma_1, \sigma_2$. That is, determine the form of $E(w) = -\log p(w|x_1, x_2, c_1, c_2, \sigma_1, \sigma_2)$.

(e) **[4 marks]** Using the negative log posterior distribution, derive the MAP estimator of w by differentiating $E(w)$ and solving $\frac{dE}{dw} = 0$ for w.

3. **Regression [14 marks]**

Suppose you are given N real-valued pairs $\{(\vec{x}_i, y_i)\}_{i=1}^N$ where the inputs are D-dimensional $\vec{x}_i \in I\!R^D$, and the output is real-valued $y_i \in I\!R$. Further you are told that a good model to explain the relation between inputs and outputs is given by

$$y_i \;=\; \exp\left(\vec{x}_i^{\,T}\vec{w}\right) + \eta_i \;,$$

where $\vec{w} \in I\!R^D$ is the vector of unknowns, and the noise η_i is an independent, Gaussian distributed random variable with mean zero and variance σ^2.

(a) **[2 marks]** Specify the mathematical form of $p(y_i \mid \vec{x}_i, \vec{w})$.

(b) **[3 marks]** Derive the form of the negative log likelihood, i.e., $E(\vec{w}) = -\log p(\{y_i\}_{i=1}^N \mid \{\vec{x}_i\}_{i=1}^N, \vec{w})$. Simplify and collect terms where possible.

(c) **[5 marks]** Derive the system of equations to be solved for the Maximum Likelihood estimate of $\vec{w}$. Note, you do not need to solve the system of equations, you simply need to derive it's form. *Hint:* Recall the following derivatives from Calculus

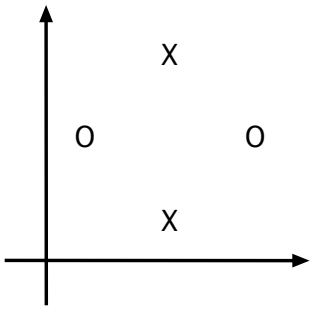
$$\frac{d}{dx}\exp(x) = \exp(x) \quad , \quad \frac{d}{dx}f(g(x)) = \frac{df}{dg}\frac{dg}{dx}$$

where $f(\cdot)$ and $g(\cdot)$ are continuous, differentiable functions.

(d) **[2 marks]** The system of equations derived above cannot be readily solved in general. What algorithm might be used instead to find the Maximum Likelihood estimate of $\vec{w}$? Explain briefly (in a sentence or two) how this algorithm works.

4. **Classification [12 marks]**

For the four 2D training data points shown below, specify whether or not each of the classification methods below, when trained appropriately, will produce zero errors when tested on the training data. In each case, briefly justify your answer.



(a) Logistic Regression

(b) Gaussian Class-Conditional Models

(c) 1-Nearest Neighbour Classification

(d) 3-Nearest Neighbour Classification

5. **Class Conditional Models [7 marks]**

Suppose we wish to use a Gaussian Class-Conditional Model to cope with an arbitrary number of classes. Denote the M classes by $c_1, c_2, \ldots, c_M$, and their means and covariances by μ_i and Σ_i for $i = 1, \ldots, M$. Let the inputs be D-dimensional real-valued vectors, $\vec{x}$, so the class-conditional distribution is

$$p(\vec{x}|C = c_i) \quad = \quad G(\vec{x}|\mu_i, \Sigma_i)$$

where $G(\vec{x}|\mu_i, \Sigma_i)$ is the Gaussian probability density function evaluated at $\vec{x}$ with mean μ_i and covariance matrix Σ_i.

(a) **[1 mark]** Is this method for classification a discriminative or generative method?

(b) **[6 marks]** Derive the posterior probability of class i given an input $\vec{x}$. That is, how do you compute $p(C = c_i|\vec{x})$. The resulting formula should be written using *only* the class-conditional model $p(\vec{x}|C = c_j)$ and the class prior distribution $p(C = c_j)$ for $j = 1, \ldots, M$.

END OF EXAM