University of Toronto at Scarborough
Department of Computer and Mathematical Sciences

CSCC11: Machine Learning and Data Mining

Midterm exam
Fall 2014

Duration: 50 minutes
No aids allowed

There are 8 pages total (including this page)

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	_____	/ 22 marks
3	_____	/ 16 marks
Total	_____	/ 53 marks

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

(a) **[2 marks]** *Bayes' Rule* can be derived by using the product rule of probability. Define the product rule and derive Bayes' Rule.

(b) **[3 marks]** Use Bayes' rule to define the posterior probability of some parameters, θ, given some data, $\mathcal{D}$. Each term (there should be three) in the resulting expression has a specific name, identify them and give the name.

(c) **[3 marks]** There are multiple ways to estimate parameters θ given data, $\mathcal{D}$. Define mathematically (that is, write equations) the Maximum Likelihood (ML) estimate, θ_{ML}, the Maximum a Posteriori (MAP) estimate θ_{MAP} and the Bayes' estimate, θ_{Bayes}. (*Hint:* The equations should be written in terms of probability distributions involving θ and $\mathcal{D}$.)

(d) **[2 marks]** The decision boundaries for Gaussian class conditional models, when there are two classes, have what mathematical form (what degree of polynomial)? If both classes have the same covariance matrix, does this change and, if so, how?

(e) **[3 marks]** Define over-fitting, explain how one can identify it and provide a strategy that might be used to prevent it.

(f) **[2 marks]** What is the key assumption used in Naïve Bayes'? Why is it used?

2. **Basis Function Regression [22 marks]** Assume you have some training data $\{x_i, y_i\}_{i=1}^{N}$, and you wish to fit a model using basis function regression, i.e.,

$$f(x_i) \;=\; \sum_{k=1}^{K} w_k b_k(x_i) + n_i \;.$$ (1)

where n_i is the Gaussian noise for the ith data point. Assume that the noise for each point is not the same, that is $n_i \sim \mathcal{N}(0, s_i)$ where s_i is the variance of the ith point.

(a) **[1 mark]** Re-define $f(x_i)$ using vector notation (with matrix-vector algebra instead of the sum) where the unknown weights are written as $\mathbf{w} = [w_1, \ldots, w_K]^T$. Clearly define any additional notation you use.

(b) **[5 marks]** Given the unknowns, $\mathbf{w} = [w_1, \ldots, w_K]^T$ and $s_{1:N} = \{s_i\}_{i=1}^{N}$, derive a mathematical expression for the negative log-likelihood, $-\log p(y_{1:N} \mid x_{1:N}, s_{1:N}, \mathbf{w})$, where $x_{1:N} \equiv \{x_i\}_{i=1}^{N}$, and $y_{1:N} \equiv \{y_i\}_{i=1}^{N}$. Include all terms and constants. (*Hint:* Remember that for $a_i > 0$, $\log(\prod_i a_i) = \sum_i \log a_i$.)

(c) **[8 marks]** Derive the gradients of $L(\mathbf{w}, s_{1:N}) = -\log p(y_{1:N} \mid x_{1:N}, \mathbf{w}, s_{1:N})$, with respect to the weight vector $\mathbf{w}$, $\frac{\partial L}{\partial \mathbf{w}}$, and a variance parameter s_i, $\frac{\partial L}{\partial s_i}$.

(d) **[4 marks]** Assuming the variances $s_{1:N}$ are known, derive the form of the optimal ML solution for $\mathbf{w}$ using a pseudo-inverse.

(e) **[4 marks]** Assuming the weights $\mathbf{w}$ are known, derive the form of the optimal ML solution for s_i.

3. **Classification [16 marks]** Pictured on this page and the next are two datasets for classification; there are two classes, the input features are 2D, and the classes are depicted with different symbols. For each of the listed classification methods (K-Nearest Neighbours (with $K = 1$ and $K = 5$), Gaussian Class Conditional and Logistic Regression) draw on the corresponding figure roughly what you expect a decision boundary for the method to look like on the provided data. Give a brief explanation (one or two sentences) in each case and say whether you think it is doing well or not and why.
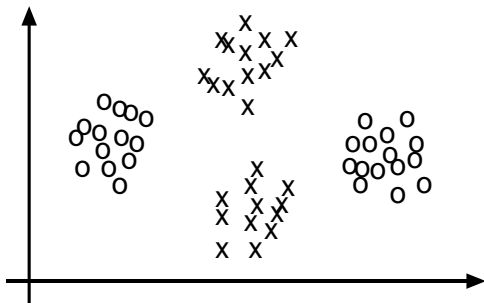
(a) **Logistic Regression:**
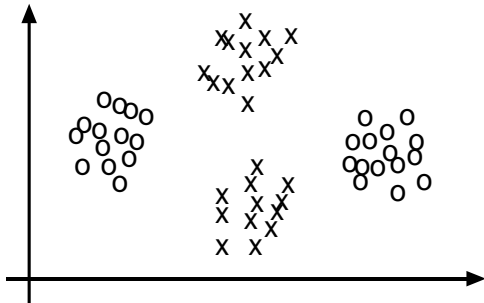
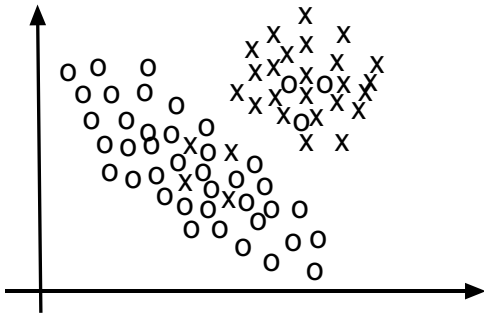(b) **K-Nearest Neighbours ($K = 1$):**
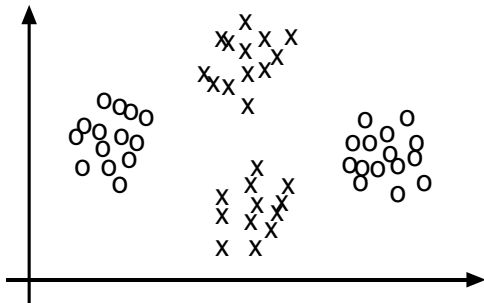
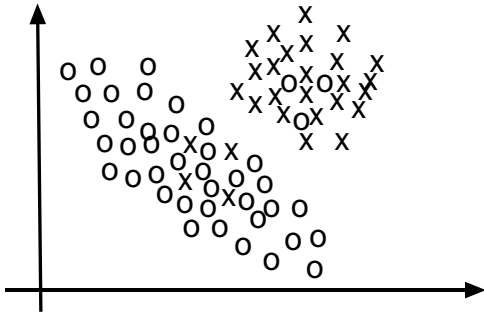(c) **K-Nearest Neighbours ($K = 5$):**
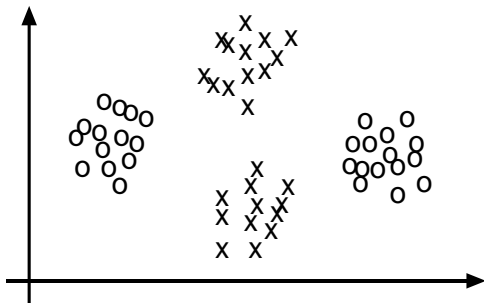
(d) **Gaussian Class Conditional:**
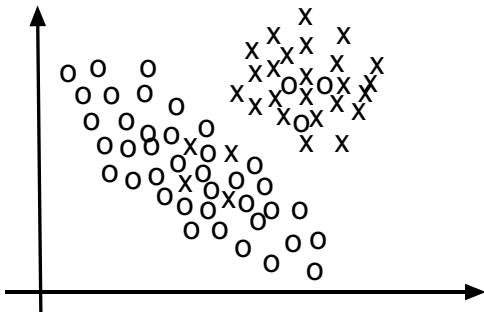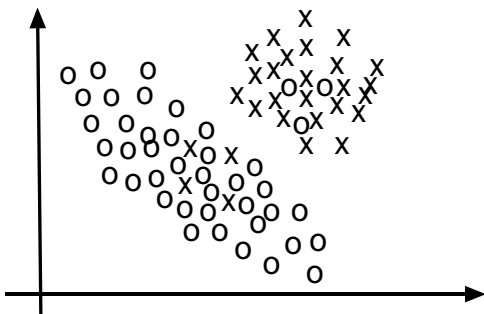
(a) **Logistic Regression:**

(b) **K-Nearest Neighbours ($K = 1$):**

(c) **K-Nearest Neighbours ($K = 5$):**

(d) **Gaussian Class Conditional:**

END OF EXAM