Quiz 1
October 24, 2016

Time: 50 minutes; Total Marks: 45
One double-sided 8.5" x 11" cheat sheet allowed

This test contains 3 questions and 5 pages

NAME: _____

STUDENT NUMBER: _____

Question	Marks	Time budget
1	/24	25 min
2	/12	10 min
3	/9	10 min

1. (24 marks) True or False questions. **Provide a short explanation.**

   (a) True or False. If a parameter μ maximizes the likelihood for a training set $\mathcal{D}$, μ also maximizes the log likelihood for $\mathcal{D}$.

   → ~~log~~ ln ↑

   log is monotonic increasing

   implies

   $\arg\max_n f(n) \rightarrow \arg\max_n \log f(n)$

   (b) True or False. The prior probability that a sample is in class k, $P(\mathcal{C}_k)$, must be no greater than 1: i.e. $P(\mathcal{C}_k) \leq 1$.

   discrete possible outputs

   all probabilities are non-negative and sum to 1

   (c) True or False. The perceptron criterion for training a classifier is equal to the number of mis-classified training examples.

   perceptron criterion:

   $$E_p(w) = -\sum_{n \in M} w^T \phi(x_n) t_n$$

   M: mis-classified examples

(d) True or False. For a fixed learning rate η, gradient descent and stochastic gradient descent will always obtain the same solution when training logistic regression.

Gradient descent and stochastic gradient descent are different

Gradient descent update the parameters using all the datapoint

Stochastic gradient descent update parameters using one single datapoint.

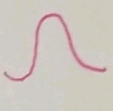
→they might obtain the same solution, but not necessarily.

(e) True or False. A neural network classifier with 1 layer of hidden units can produce non-linear decision boundaries.

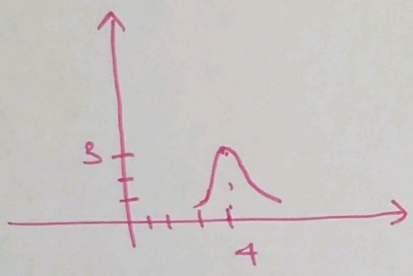
(f) True or False. The weight vector $\boldsymbol{w}$ that minimizes error in a neural network is unique.

2. (12 marks) Consider regression with a single training data point: $(x_1 = 4, t_1 = 3)$ and the basis function

$$\phi_1(x) = \exp\{-(x-4)^2\}$$ → Gaussian basis function



- Suppose we train a model with no regularization using only the basis function $\phi_1(x)$ (no bias term): $y(x) = w_1\phi_1(x)$.
  - Draw the learned function $y(x)$.
  - What would w_1 be?



$y(x) = w_1\phi_1(x) + w_0$ ($w_0 = 0$)

$y(x) = w_1\phi_1(x)$

$y(x=4) = w_1\phi_1(x=4)$

$\phi_1(x=4) = \exp\{-(4-4)^2\} = 1$

$\rightarrow y(x=4) = w_1 \times 1 = 3 \rightarrow w_1 = 3$

- Suppose we added a bias term: $y(x) = w_0 + w_1\phi_1(x)$ and trained with no regularization. What would happen?

$E(w)$ = sum of squared error + ~~regularization~~ term = 0

$E(w)$ = sum of squared error → only a single data point

$E(w) = (y(x=4) - t)^2 = (w_0 + w_1\phi(x=4) - 3)^2 =$

$(w_0 + w_1 - 3)^2$

$\frac{\nabla E}{\nabla w_0} = 2 \times (w_0 + w_1 - 3) \rightarrow w_0 + w_1 - 3 = 0$

$\frac{\nabla E}{\nabla w_1} = 2 \times (w_0 + w_1 - 3) \rightarrow w_0 + w_1 - 3 = 0$

no unique solution

- Suppose we added a bias term: $y(x) = w_0 + w_1\phi_1(x)$ and trained with regularization only on w_1. What would happen?

$E(w)$ = sum of squared error + regularization term on w_1

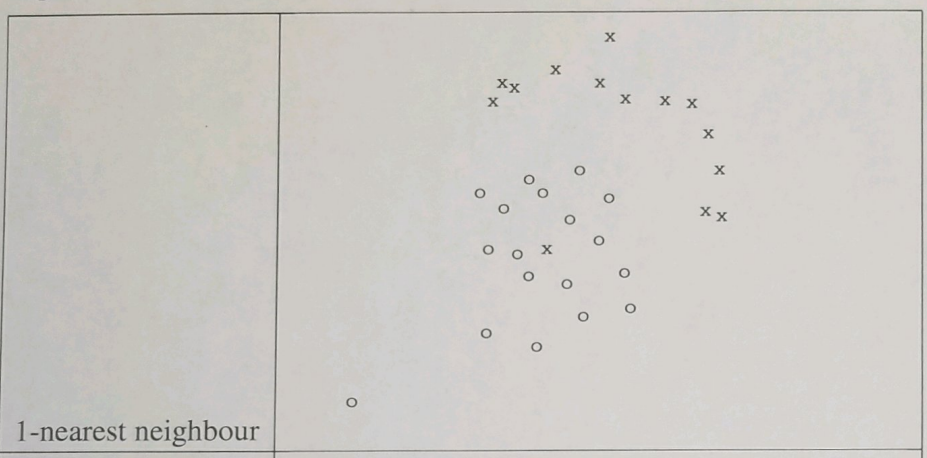
$E(w) = \frac{1}{2}(y(x=4) - 3)^2 + \frac{\lambda}{2}w_1^2 = \frac{1}{2}(w_0 + w_1 - 3)^2 + \frac{\lambda}{2}w_1^2$

$\frac{\nabla E}{\nabla w_1} = \frac{1}{2} \times 2 \times (w_0 + w_1 - 3) + \frac{\lambda}{2} \times 2 \times w_1 = 3 \rightarrow w_0 + (1+\lambda)w_1 = 3$

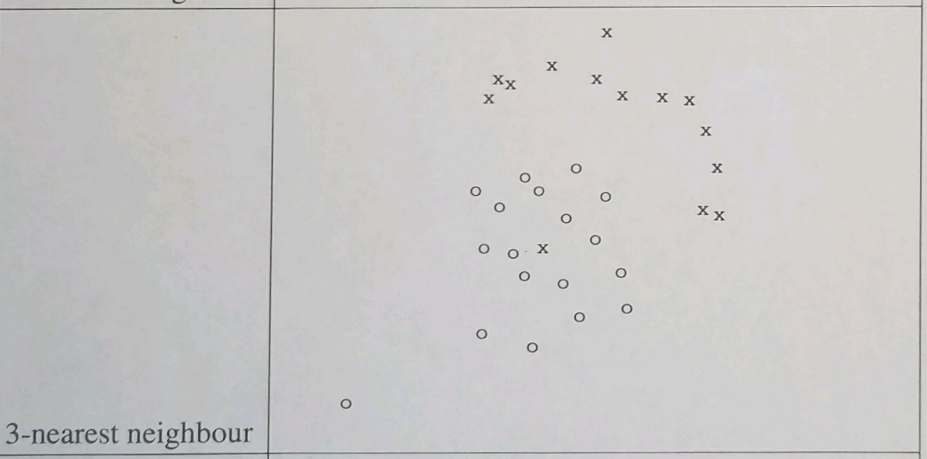
$\frac{\nabla E}{\nabla w_0} = \frac{1}{2} \times 2 \times (w_0 + w_1 - 3) = 0 \rightarrow w_0 + w_1 = 3$

$\rightarrow$ $w_1 = 0$, $w_0 = 3$

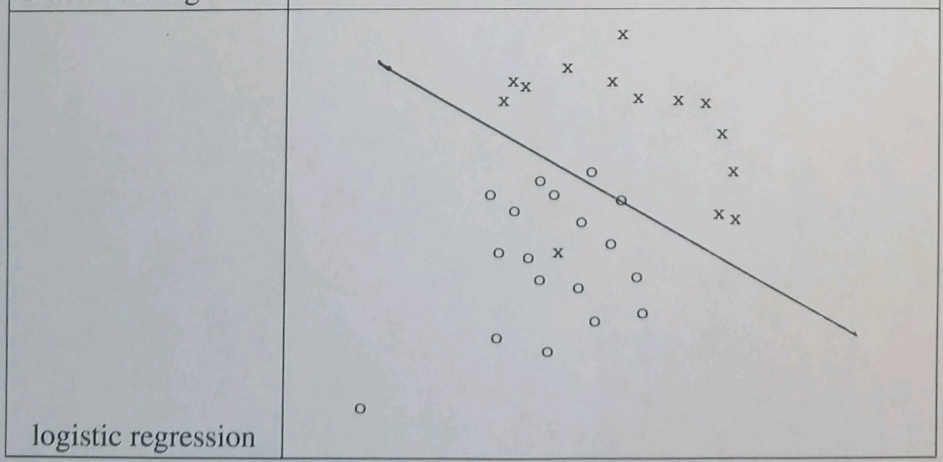
3. (9 marks) Consider the training set below for two-class classification. Draw the approximate decision regions when using **1-nearest neighbour**, **3-nearest neighbour**, and **logistic regression**. Please notice the "x" in the middle of the "o" points.



1-nearest neighbour



3-nearest neighbour



logistic regression