ECE/CS/ME 539 Introduction to Artificial Neural Networks

Homework 03

The homework is graded on *completion only* scale. If an answer to the problem exists, full credit will be given regardless the correctness of the answer. No answer will receive no credit though.

Coverage: Data Preprocessing, Performance metrics, PCA

1. (2 pts) (*Confusion Matrix*) A new test is developed for detecting the Covid-19 virus. After experiment, the confusion matrix of this test is reported below:

		Predicted labels	
		P	**N**
Actual labels	**P**	**15**	**5**
	N	**4**	**16**

How many positive (P) results this new test reports? 19

What percentage of true Covid-19 cases are detected among all true Covid-19 cases in this test? 78.947%

What is the false positive rate, defined as fraction of cases that are reported as positive but are negative, among all tests? 4/19

2. (4 pts) (*Performance metrics*) Consider 10 feature-label pairs $\{x(k), t(k); 1 \le k \le 10\}$. Assume that $t(k) \in \{0, 1\}$. The posterior probability $P(t(k) = 1|x(k))$ is given in the 2nd row of the table below and the corresponding ground truth label is given in the third row. Here a class label = 0 means Negative, and a class label = 1 means positive.

Index k	1	2	3	4	5	6	7	8	9	10
$P(t(k) = 1\|x(k))$	0.05	0.15	0.40	0.55	0.25	0.45	0.48	0.62	0.67	0.75
True Label	0	0	0	0	1	1	1	1	1	1
Predicted label $y(k)$	0	0	1	1	0	1	1	1	1	1

Given a threshold b, we set the predicted label $y(k) = 0$ if $P(t(k) = 1|x(k)) \le b$; and = 1 otherwise.

(a) If $b = 0.3$, fill in the predicted label in the 4th row of above table.
Answer: hw03P2.mlx, hw03.ipynb

Predicted label $y(k)$	0	0	1	1	0	1	1	1	1	1

predicted

actual	0	1
0	2	2
1	1	5

(b) Compute the confusion matrix **C** with $b = 0.3$
(c) With $b = 0.3$, compute the following quantities: sensitivity (sen), specificity (spe), Pr. False Alarm (pfa), Pr. Miss (pmiss), precision (pre), recall and classification rate (crate).
(d) For the value of threshold b varying from 0 to 1, compute the list of distinct pairs of (TPR, FPR) and then plot the ROC curve and calculate the area under the (ROC) curve (AUC).

3. (4 pts) (*PCA*) Download the dataset **`mnist_test.csv`**. This dataset consists of 28 by 28 gray level images of hand-written numerals from 0 to 9. The first column contains the labels {0, 1, …, 9}. The remaining columns contain the image of hand-written numerals in the form of row vectors.

(a) Number of samples N = 16000. Feature dimension = 784.

(b) Visualize the first 20 rows (samples). Each should be displayed as a 28 by 28 image. If you are using python, you may refer to this blog post:
How to Load and Plot the MNIST dataset in Python?

(c) Denote the N by d feature matrix $\mathbf{X}$. Perform SVD of $\mathbf{X}$. Design the singular values as a vector $\mathbf{s}$. plot $\log_{10}(\mathbf{s})$ over the range 1 to d.

(d) Denote the first two principal components by a d by 2 matrix $\mathbf{V}$. Use the first 2 principal components, projecting each row of the $\mathbf{X}$ matrix by computing $\mathbf{Z} = \mathbf{XV}$. Each row of $\mathbf{Z}$ is a 1 by 2 vector corresponding to a point in a 2D space spanned by the two columns of $\mathbf{V}$. Give a scatter plot of these projected 2D points corresponding to numeral 1 and 8. Note that the numerals are class labels.

(e) This one is for all 10 numerals. An approximation of the original feature matrix $\mathbf{X}$ may be estimated as:

$$\hat{\mathbf{X}} = \mathbf{XVV}^T = \mathbf{ZV}^T$$

Visualize the corresponding 28 by 28 patterns of the first 20 rows of $\hat{\mathbf{X}}$ (Xhat)