MLAI4DS Mock Exam Paper

This examination paper is worth a total of 60 marks

1. Linear regression with models of the form $t_n = \mathbf{w}^T\mathbf{x}_n$, is a common technique for learning real-valued functions from data.

(a) Squared and absolute loss are defined as follows:

$$L_{squared} = (t_n - \mathbf{w}^T\mathbf{x}_n)^2, \quad L_{absolute} = |t_n - \mathbf{w}^T\mathbf{x}_n|$$

Describe, with a diagram if you like, why, when optimizing the parameters with the squared loss outliers have a larger effect than with the absolute loss.

[6 marks]

Defining e_n as t_n –w^Tx_n, for values of |e_n| < 1, the absolute loss is larger than the squared loss [2]. However, as e_n increases, the squared loss increases much faster [2]. Outliers have a high e_n and therefore a much larger influence over squared loss than in absolute loss[2]

(b) Which of the following statements is true:

A) Parameter estimation with the squared loss is not analytically tractable.

B) The squared loss is equivalent to assuming normally distributed noise.

C) The absolute loss is a popular choice for regularization.

D) The squared loss is a popular choice for regularization.

[2 marks]

B

(c) Discuss why the value of the squared loss on the training data cannot be used to choose the model complexity.

[3 marks]

When training the model we are minimizing the loss on the training data [1]. As we make models more complex, the squared loss value at the minima will always decrease [2]. Hence, training loss will always favour more complex models.

(d) For the particular model $t_n = w_1 x_n + w_2 x_n^3$, I optimize the parameters and end up with $\mathbf{w} = [2,1]^T$. What does the model predict for a test point at $x_{new} = 3$?

[2 marks]

2*3 + 3^3 = 33

(e) The radial basis function (RBF):

$$h_{n,k} = \exp\left(-\frac{\sum_{d=1}^{D}(x_{n,d} - \mu_{d,k})^2}{2s^2}\right), n = 1, \ldots, N;\ k = 1, \ldots, K$$

is a popular choice for converting the original features, $x_{n,d}$, into a new set of K features

prior to training. Assume the value of s is given. Describe a procedure for determining the center parameter $\mu_{d,k}$ and K.

[4 marks]

Any reasonable approach for choosing K centers, the answer should clearly state how to choose K [2] and what is $\mu_{d,k}$[2]. A common choice is let $\mu_{d,k}$ be a data point $x_{i,d}, i \neq n$ other than the nth data point. In this case, K = N-1.

(f) With respect to the functions they can fit, describe the difference between RBF and the basic linear model $\mathbf{w}^T\mathbf{x}_n$ with a graph.

[3 marks]

The RBF can produce a nonlinear model i.e. wiggly curve [1]. $\mathbf{w}^T\mathbf{x}_n$ produces only the linear model i.e. straight lines or hyperplanes [1]. 1 mark for readable graph.

2. Classification question

(a) Use a classification algorithm, describe what is meant by:

(i) Generalisation

[2 marks]

(ii) Over-fitting

[2 marks]

Plenty of sensible answers. Marks awarded for a description of generalization that describes the model's ability to make predictions on previously unseen data and, for overfitting, the problem of *memorizing* the training data.

(b) A classification algorithm has been used to make predictions on a test set, resulting in the following confusion matrix:

		Truth		
Prediction		**Positive**	**Negative**	**Total**
	Positive	23	5	28
	Negative	10	12	22
	Total	33	17	50

Compute the following quantities (expressing them as fractions is fine):

(i) Accuracy

[2 mark]

35/50

(ii) Sensitivity

[2 mark]

23/33

(iii) Specificity

[2 mark]

12/17

(c) Explain why it is not possible to compute the AUC from a confusion matrix.

[4 marks]

AUC is the area under the ROC curve [1]. The ROC curve is created by varying the threshold at which the algorithm calls something as belonging to the positive class [1]. A confusion matrix only gives us the performance at one threshold [1].

(d) Two binary classifiers are used to make predictions for the same set of six test points. These predictions are given below, along with the true labels. Compute its area under the curve (AUC) in each case.

Classifier 1		Classifier 2	
Predicted probability of class 1 (Score of class 1)	True class	Predicted probability of class 1 (Score of class 1)	True class
1	1	0.8	1
0.8	1	0.8	0
0.6	1	0.6	1
0.4	0	0.6	0
0.2	0	0.2	1
0.0	0	0.2	0

[4 marks]

AUC for classifier 1 is 1 [2] and 0.5 for classifier 2 [2].

(e) Explain how the SVM can be extended via the kernel trick to perform non-linear classification.

[2 marks]

In the SVM objective function, the data only appear in the form of inner products [1]. Inner products in the original space can be replaced by inner products in another feature space via kernel functions [1].

3. Unsupervised learning

(a) Provide pseudo code for K-means (assume that the number of clusters is provided).

[5 marks]

Given: Number of clusters, K

2. For each cluster k = 1...K:

3. For each object n = 1...N: [1]

4. Compute the distance between object n and cluster k [1]

5. Assign object n to the cluster corresponding to the smallest distance [1]

6. Update the mean of each cluster [1]

7. If assignments have changed, return to 2. Else stop. [1]

(b) Is the total Euclidean distance between data points and their cluster centers a good criterion to select number of clusters in K-means? Why?

[2 marks]

No [1], large number of clusters will lead to smaller and better Euclidean distance [1]

(c) Gaussian mixture models can be fitted to data using the expectation maximization (EM) algorithm. The EM algorithm has two steps: E-step and M-step. Describe what parameters are being estimated in each step in Gaussian mixture models.

[4 marks]

In the Estep, the estimated parameters are the expected assignment probabilities for each data point to each Gaussian component [2]. In the Mstep, the mean and covariance of Gaussian components and the mixing coefficients [2].

(e) Describe three key differences between K-means and Gaussian mixture models

[4 marks]

K-means use hard assignment [1]. Gaussian mixture models estimate the probability of assignment i.e. soft assignment [1]. GMM estimate both cluster center [2] and covariance [1].

(e) K-means often converges to a local optimal solution. Describe a simple process for overcoming the local optimality of K-means.

[3 marks]

For fixed K, perform multiple re-starts, and evaluate the total distance of points from their cluster mean. Keep the solution with the smallest distance.

(g) Describe two situations (with justification) where you might choose a mixture model over K-means.

[2 marks]

When using a data-type for which it is not obvious how to compute a distance (but for which we can compute a likelihood), or when we have data for which we don't want to be limited to isotropic clusters.