ELL409: Machine Intelligence and Learning

Minor Test II, Form: A

Maximum marks: 20

(Answer all questions on this question paper. Use the answer script only for working; it will not be graded. Read all section-specific instructions carefully.)

Name: _____

Entry Number: _____

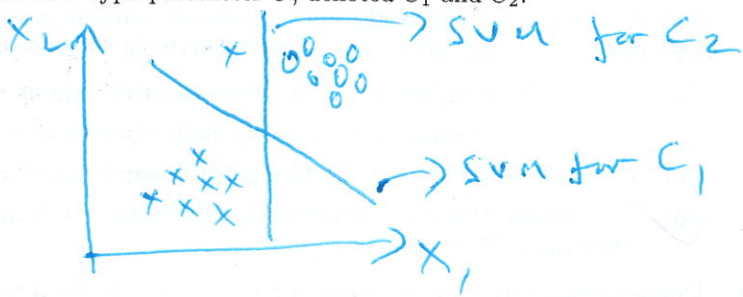
Section 1. Multiple choice questions

Instructions: Each question may have any number of correct choices. Clearly mark (tick) all choices you believe to be correct (1 mark for each correct choice, −0.5 for each incorrect choice).

1. You are training an RBF SVM with the following parameters: C (slack penalty) and σ (where σ^2 is the variance of the RBF kernel). How should you tweak the parameters when you find the model to be underfitting?

   (a) Increase C and/or reduce σ
   (b) Reduce C and/or increase σ
   (c) Reduce C and/or reduce σ
   (d) Increase C and/or increase σ
   (e) Increase C only (σ has no predictable effect on underfitting)

2. Consider the below figure showing some two-class training data, and linear SVM decision boundaries learnt for two distinct values of the hyperparameter C, denoted C_1 and C_2.



   Which of the following are true?

   (a) $C_1 > C_2$.
   (b) $C_1 < C_2$.
   (c) C_1 leads to a higher sum total of slack (ξ_i) values.
   (d) C_2 leads to a higher sum total of slack (ξ_i) values.
   (e) C_1 corresponds to the model with higher variance, of the two.
   (f) C_2 corresponds to the model with lower bias, of the two.

3. For an RBF SVM with a particular pair of randomly chosen values of the hyperparameters C and σ (where σ^2 is the variance of the RBF kernel), which of the following tests can be taken to indicate that the chosen values likely correspond to underfitting?

   (a) When I increase C the training and validation accuracies both increase.

   (b) When I decrease C the training accuracy reduces, but validation accuracy increases.

   (c) When I decrease C the training and validation accuracies are both reduced.

   (d) When I decrease σ the number of support vectors increases.

   (e) When I decrease σ the training accuracy reduces, but validation accuracy increases.

   (f) When I increase σ the training accuracy increases, but validation accuracy reduces.

4. Consider the following possible choices of error function in training a logistic regression model for classification: cross-entropy error (I), classification error (II), and sum-of-squares error (III). Which of the following are true?

   (a) (II) is problematic because it's non-differentiable, but either of (I) or (III) should give the same result.

   (b) Any of the three could be easily used for gradient descent, but (I) is preferred because it corresponds to maximising the likelihood of the data.

   (c) (II) is problematic because it's non-differentiable; (III) is preferred to (I) because the latter corresponds to an inappropriate noise model.

   (d) (II) is problematic because it's non-differentiable; (I) is preferred to (III) because the former corresponds to maximising the likelihood of the data.

   (e) Any of the three could be easily used for gradient descent, but (III) is preferred because it corresponds to maximising the likelihood of the data.

5. You are given a labeled binary classification data set with N data points and D features. Suppose that $N < D$. Which of the following kinds of feature transformation do you think would generally make sense in such a setting, prior to or as part of training a classifier?

   (a) Using Gaussian/sigmoidal basis functions to obtain more fine-grained or localised features

   (b) Using a polynomial kernel SVM to implicitly map to a higher-dimensional feature space

   (c) Using an RBF kernel SVM to implicitly map to an infinite-dimensional feature space

   (d) No feature transformation should be necessary, a simple linear classifier trained on the raw features is likely to work

6. Suppose your model is demonstrating high bias across different training sets. Which of the following are valid ways to try and reduce the bias?

   (a) Decrease the amount of training data in each training set.

   (b) Improve the optimisation algorithm being used for error minimisation.

   (c) Increase the model complexity.

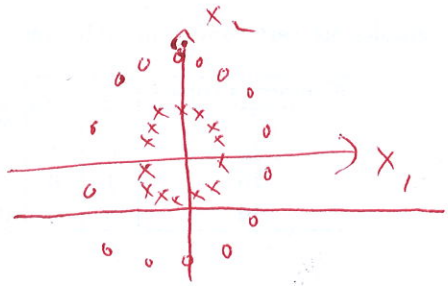
   (d) Increase the noise in the training data.

Section 2. Numerical/Short-answer questions

Instructions: Please write *only the final answers* on this question paper, in the space provided for each item. The provided answer script should be used for all working, but will not be graded. However, in case of any doubt regarding your answers, we may refer to the answer script to check your working. So please try to write out your working as clearly as possible.

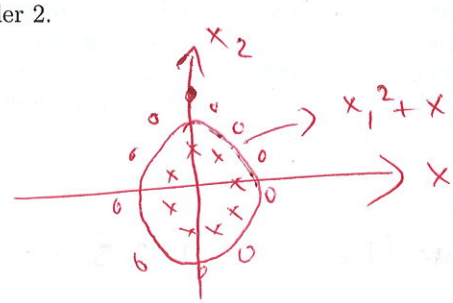
7. Suppose we have a two-class data set in 2-D space, generated as follows: positive samples taken from points on the curve $x_1^2 + x_2^2 = 5$, and negative samples taken from points on the curve $x_1^2 + x_2^2 = 10$. The number of samples in both classes are equal. Show visually, in the input feature space, the kind of decision boundary that would be obtained by training an SVM (with suitable hyperparameter tuning) with each of the following choices of kernel. Also mention in each case what you think would be (at least roughly) the equation defining the decision boundary in input space.

   (a) Linear kernel. **[0.5]**



e.g. $x_2 = -\sqrt{5}$ (others also feasible)

   (b) Polynomial kernel of order 2. **[1]**



$x_1^2 + x_2^2 = \left(\frac{\sqrt{5}+\sqrt{10}}{2}\right)^2$ (7.5 also OK)

   (c) RBF kernel. **[1]**

same as above

8. You are given a small data set with just 4 points in 2-D space. Two positive examples, with coordinates $(1, 4)$ and $(2, 3)$; and two negative examples, with coordinates $(4, 5)$ and $(5, 6)$.

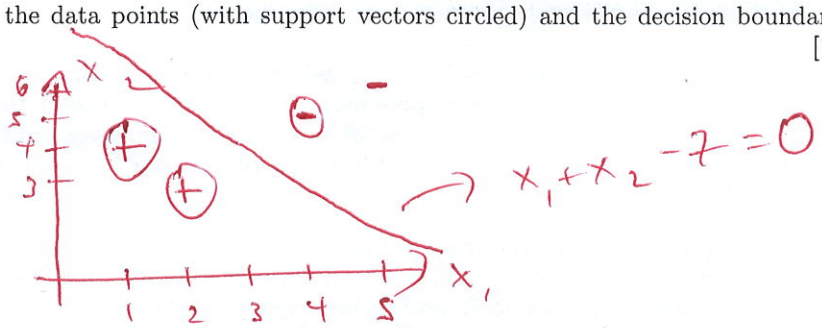
   (a) What will be the weight vector $\mathbf{w}$ corresponding to the maximum-margin decision boundary learnt by a linear hard-margin SVM on this data set? **[1.5]**

$\underline{w} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

   (b) What will be the bias term b corresponding to the same decision boundary? **[0.5]**

$b = -7$

(c) Draw a plot showing the data points (with support vectors circled) and the decision boundary learnt. [1]



9. Here we will look at methods for selecting input features for a logistic regression model

$$P(t = 1|\mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1 x_1 + w_2 x_2).$$

The available training examples are very simple, involving only binary valued inputs:

Number of copies	x_1	x_2	t
5	1	1	0
10	0	1	1
10	1	0	1
10	0	0	0

So, for example, there are 5 copies of $\mathbf{x} = (1, 1)^{\mathrm{T}}$ in the training set, all labeled $t = 0$. The correct label is actually a deterministic function of the two features: $t = 0$ if $x_1 = x_2$ and $t = 1$ otherwise. We define greedy selection in this context as follows: we start with no features (train only with w_0) and successively try to add new features provided that each addition strictly reduces the training error (cross-entropy). We use no other stopping criterion.

(a) Which feature would greedy selection add first: x_1, x_2, or neither? Briefly mention the reason for your answer. [1]

Neither; 20/35 classified correctly with a constant model ($t=1$), no improvement for either x_1 or x_2

(b) What is the minimum *classification error* on the training examples that we could achieve by including both x_1 and x_2 in the logistic regression model? [1]

$\frac{5}{35} = 14.3\%$

(c) Suppose we define another possible feature to include, a function of x_1 and x_2. Which of the following features, if any, would permit us to correctly classify all the training examples when used in combination with x_1 and x_2 in the logistic regression model: $x_1 - x_2$, $x_1 x_2$, x_2^2? (Specify all that apply.) [1.5]

$x_1 x_2$

(d) Suppose we add the set of features identified in response to part (c) to our feature set, and then use the greedy feature selection method as above. Which feature would it now add first? Briefly mention why. [1]

$x_1 x_2$: improves correctly classified pts. to 25/35.