Major Exam

Total marks: 30 pts

Total time: 1 hr 40 mins

HONOUR CODE
As a student of IIT Delhi, you will not give or receive aid in examinations. You will do your share and take an active part in seeing to it that others as well as yourself uphold the spirit and letter of the Honour Code.

Instructions:

- **Write your name and roll number on answer script**
- With the exception of Question 1, all your answers must be clearly motivated! *A correct answer without a proper motivation will score zero points*!
- Vectors are denoted with a single underline $\underline{a}$, and matrices by double underline, $\underline{\underline{A}}$. Scalars appear without any underline. Please follow this rule through your answer book.

1. **[8 pts]** Answer `True` or `False`.
   Each correct answer scores 1 point, each incorrect answer scores -1 point and each missing answer scores 0 point.

T (a) In ensemble methods, averaging over parameters (instead of predictions) can not be a valid method?

T (b) The `flatten` function is used to transform the activations of the convolution layers, which have channel, height, and width dimensions, into flat vectors that are input into the fully connected layers

F (c) The parameters of hard-margin support vector regression have a closed form solution

F (d) Neural Networks are non-parametric methods

F (e) To talk about the bias-variance trade-off only has meaning when learning by minimizing the mean squared error cost function.

(f) Please select all that apply about k-NN in the following options. Assume a point can be its own neighbor.

0.5 ✓ (A) k-NN works great with a small amount of data, but struggles when the amount of data becomes large.

-0.5 for incorrect (B) k-NN is sensitive to outliers; therefore, in general we decrease k to avoid overfitting.

(C) k-NN can be applied to classification problems but not regression problems.

0.5 ✓ (D) We can always achieve zero training error (perfect classification) with k-NN, but it may not generalize well in testing.

F (g) Logistic regression and quadratic discriminative analysis (QDA) will always produce the same decision boundary for binary classification problems

T (h) k-means clustering algorithm can be considered a special case of Gaussian Mixture Model

2. **[8 pts]** A company is working on a machine learning problem with $p = 5000$ input variables $\underline{x} = \{x_1, \ldots, x_{5000}\}^T$ with the aim to predict a single binary output $y \in \{1, -1\}$. They have collected $N = 200$ datapoints with equally many examples from each class.

(a) **[3 pts]** From previous experience, most of the features are useless for predicting y. To build a model, the company has executed the following steps:

i) Compute the linear correlation between each input and the output and keep only the 100 most correlated input variables (positive or negative correlation).

ii) Learn a logistic regression model.

iii) Evaluate the performance of the model by running 10-fold cross-validation to estimate the accuracy on new, unseen data.

The cross-validation indicated an accuracy of over 95% for the model. However, when they tried it in production, the performance was much worse.

Is the problem too hard or did the company do anything wrong? Give advice as to how the company should act.

(b) **[4 pts]** The company also tried LASSO. They learned the parameters of a logistic regression model by minimizing the cost function

$$J(\underline{\theta}) = \sum_{i=1}^{200} \log\left(1 + e^{-y_i \underline{\theta}^T \underline{x}_i}\right) + \frac{1}{\lambda} \sum_{j=1}^{5000} |\theta_j|$$

To learn the regularization parameter λ, they have produced the following graph.

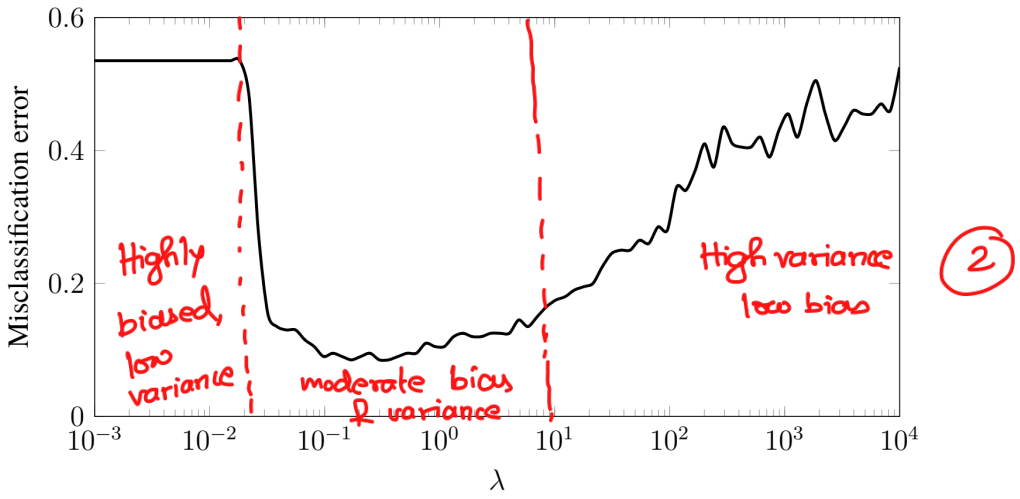

Figure 1: Misclassification error estimated using 10-fold cross-validation for different values of the regularization parameter λ

Explain why the graph looks the way it does. Specifically, explain the shape in terms of the bias-variance trade-off and which term dominates where. Explain also what bias and variance mean in terms of model complexity and how this relates to λ.

(c) **[1 pt]** The left part of the graph is completely flat. Why?

① To clear out the gradients of all parameters that optimizer is tracking

3. **[2 pts]** What is the purpose of `zero_grad` method in PyTorch training of neural networks? What happens if the method is not called?

① Gradients from previous minibatches would add up leading to incorrect gradient computation

4. **[2 pts]** In the linearly separable case of SVM if one of the training samples is removed, will the decision boundary shift toward the point removed or shift away from the point remove or remain the same? Justify your answer.

If training sample == support vector, then yes
" " ≠ support vector, then no } ②

5. **[6 pts]** In the lecture, we covered the soft-margin SVM. over the training set $\{(\underline{x}_1, y_1), (\underline{x}_2, y_2),..., (\underline{x}_n, y_n)\}$.

Here, we are interested in a modified version of the soft-margin SVM where we have a custom margin ξ_i for each of the n data points. In the standard soft-margin SVM, we pay a penalty of ξ_i for each of the data point. In practice, we might not want to treat each training point equally – for example, we might know that some data points are more important than the others.

We formally define the following optimization problem:

$$\min_{\underline{\theta},b,\{\xi_i\}} \frac{1}{2}\|\underline{\theta}\|_2^2 + C\sum_{i=1}^{n}\phi_i\xi_i$$
$$\text{s.t.} \quad y_i(\underline{\theta}^T\underline{x}_i - b) \geq 1 - \xi_i \quad \forall i$$
$$\xi_i \geq 0 \quad \forall i$$

$\xi_i \geq 1 - y_i(\underline{\theta}^T\underline{x}_i - b)$

$\xi_i \geq 0$ Since $\sum \xi_i$ is being minimized therefore

Note that the only difference is that we have a custom weighting factor $\phi_i > 0$ for each of the slack variables ξ_i in the objective function. These ϕ_i are some constants given by the prior knowledge, and thus they can be treated as known constants in the optimization problem. Intuitively, this formulation weighs each of the violations (ξ_i) differently according to the prior knowledge (ϕ_i).

(a) **[2 pts]** For the standard soft-margin SVM, we have shown that the constrained optimization problem is equal to the following unconstrained optimization problem:

$$\min_{\underline{\theta},b} \frac{1}{2}\|\underline{\theta}\|_2^2 + C\sum_{i=1}^{n}\max(0, 1 - y_i(\underline{\theta}^T\underline{x}_i - b))$$

$\xi_i \geq \max(0, 1 - y_i(\underline{\theta}^T\underline{x}_i - b))$

Motivation has partial marks

What's the corresponding unconstrained optimization problem for the SVM with custom margins?

(b) **[4 pts]** As seen in lecture, the dual form of the standard soft-margin SVM is:

$$\min_{\underline{\alpha}} \frac{1}{2}\underline{\alpha}^T\underline{\underline{K}}\underline{\alpha} - \underline{\alpha}^T\underline{1}$$
$$\text{s.t.} \sum_{i=1}^{n}\alpha_i y_i = 0$$
$$0 \leq \alpha_i \leq C \quad i = 1, ..., n$$

where $\underline{\underline{K}} = (\text{diag}(\underline{y}))\underline{\underline{X}}\,\underline{\underline{X}}^T(\text{diag}(\underline{y}))$.

What is the dual form of the SVM with custom margins? To start, you are provided with the Lagrangian, which is given by

Step marking is followed!

$$\mathcal{L}(\underline{\theta}, b, \xi, \alpha, \gamma) = \frac{1}{2}\|\underline{\theta}\|_2^2 + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i y_i(\underline{\theta}^T\underline{x}_i - b) + \sum_{i=1}^{n}(C\phi_i - \alpha_i - \gamma_i)\xi_i$$

The derivation is very similar to one done in lecture.
Only difference will be in $0 \le \alpha_i \le \phi_i C$

6. **[4 pts]** Recall that in k-means clustering, we attempt to minimize the objective → J

$$\min_{r_{ik},\underline{\mu}_k} \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}\left\|\underline{x}_i - \underline{\mu}_k\right\|_2^2$$

where it turned out that: if the assignment r_{ik} is known for each $\underline{x}_i$, the mean of the kth cluster is given by

$$\underline{\mu}_k = \frac{1}{\sum_{i=1}^{N} r_{ik}} \sum_{i=1}^{N} r_{ik}\,\underline{x}_i$$

The samples are $\{\underline{x}_1, \ldots, \underline{x}_N\}$, where $\underline{x}_i \in \mathbb{R}^d$.

(a) **[1 pt]** What is the minimum value of the objective when $K = N$ (the number of clusters equals the number of sample points)? Each datapt will be a cluster ⇒ J = 0 — (1)

(b) **[3 pts]** Suppose we add a regularization term to the above objective, as follows:

$$\min_{r_{ik},\underline{\mu}_k} \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}\left\|\underline{x}_i - \underline{\mu}_k\right\|_2^2 + \lambda\left\|\underline{\mu}_k\right\|_2^2$$

.

Determine the expression for $\underline{\mu}_k$, assuming known r_{ik} from a previous assignment step.

$$J(\underline{\mu}) = \sum_{i=1}^{N}\sum_{k=1}^{N} r_{ik}\,(\underline{x}_i - \underline{\mu}_k)^T(\underline{x}_i - \underline{\mu}_k) + \lambda\,\underline{\mu}_k^T\underline{\mu}_k$$

$$\frac{\partial J}{\partial \underline{\mu}_k} = 0 \Rightarrow \frac{\partial J}{\partial \underline{\mu}_k} = -\sum_{i=1}^{N} r_{ik}\,(\underline{x}_i - \underline{\mu}_k) + \lambda\underline{\mu}_k = 0 \quad (1)$$

(1)

$$\Rightarrow \left(\sum_{i=1}^{N} r_{ik} + \lambda\right)\underline{\mu}_k = \sum_{i=1}^{N} r_{ik}\,\underline{x}_i$$

$$\Rightarrow \underline{\mu}_k = \frac{\sum_{i=1}^{N} r_{ik}\,\underline{x}_i}{\sum_{i=1}^{N} r_{ik} + \lambda} \quad (1)$$