# Indian Institute of Technology, Kharagpur

*Department of Computer Science and Engineering*

## Mid Semester Examination, Autumn 2023

Machine Learning (CS60050)

**Students**:156 **Date**: 26-Sep-2023

**Full marks**: 57 **Time**: 2:00pm–4:00pm

**Instructions:** Answer the questions in the boxes provided in this answer booklet only. You may use supplementary sheets for rough work which must be tied to this booklet.

1. (a) When the number of features is increased in a linear model, state whether the following are expected to increase or decrease with proper justifications: [4]
   i. Bias of the model
   ii. Variance of the model

**Solution:**

1. Bias: (b) Decreases
   Justification: More features means more powerful model to fit the data.
2. Variance: (a) Increases
   Justification: If the dataset consists of too many features for each data-point, the model often starts to suffer from high variance and starts to overfit.

(b) How do the relative error on the following quantities expected to change when you add regularization such as L2 regularization? State whether "increases" or "decreases". [2]
   i. Training error
   ii. True error

**Solution:**

1. Training error: increases
2. True error: decreases

(c) Consider a linear regression model with $L_1$-regularization. [2]

$$Loss(\boldsymbol{\theta}) = \frac{1}{m}\sum_{n=1}^{m}(y_n - f(\boldsymbol{x}))^2 + \lambda\boldsymbol{\theta} \quad \text{where } f(\boldsymbol{x}) = \boldsymbol{\theta}^T\boldsymbol{x}$$

How do bias and variance change as $\lambda$ is increased?

**Solution:** Bias Increases.
Variance Decreases.

(d) Suppose you have a task to classify a bird species into one of 23 classes. Your data for each bird is described by 20 numerical features. How many parameters are required if you use a linear model for this bird classification task? [2]

**Solution:** $23 \times 21$

(e) Write the expression for the hypothesis $f(\boldsymbol{x})$ and the loss function $L$ used in logistic regression. Assume that there are $N$ training examples and $d$ input features. The training set is given by $\{(\boldsymbol{x}^{(1)}, y^{(1)}), \cdots (\boldsymbol{x}^{N}, y^{(N)})\}$ and the $i$th feature of the $j$th example is given by $\boldsymbol{x}_j^{(i)}$. [3]

**Solution:**

$$f(\boldsymbol{x}) = \frac{1}{1 + e^{-\theta^T x}}$$

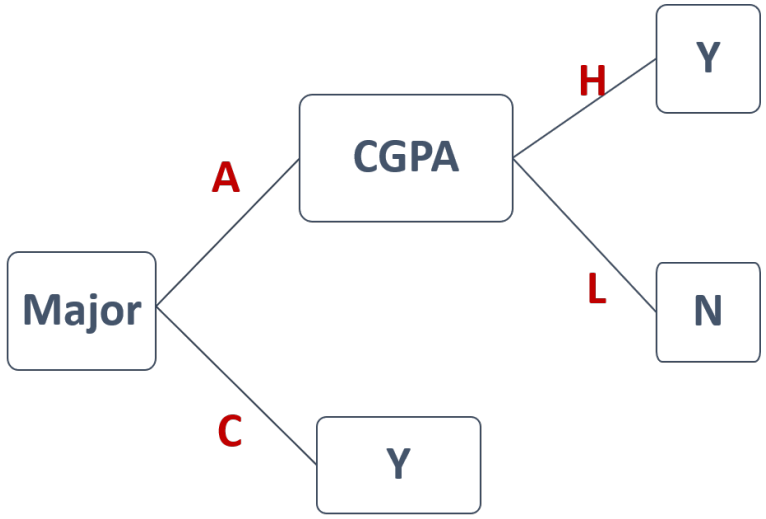
$$L = -y\log(f_\theta(x)) - (1-y)\log(1 - f_\theta(x))$$

2. (a) Use the following dataset to learn a decision tree for predicting if a candidate is short-listed for a job (Y) or not (N), based on their Major (A or C), CGPA (High or Low), and whether they did a Project (Y or N). [4]

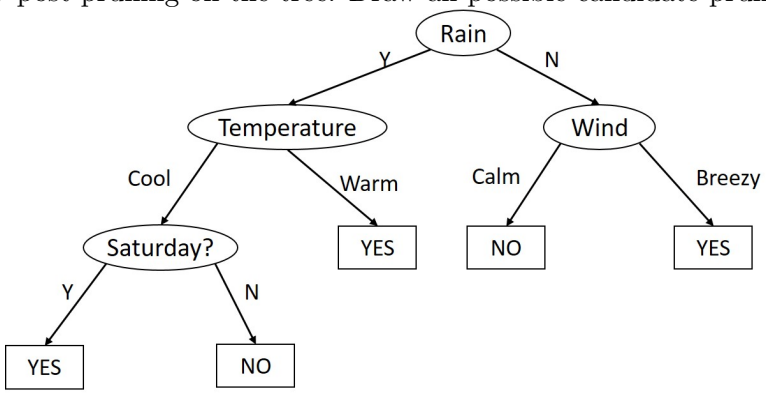
Draw the full decision tree that would be learned for this data using the Information Gain heuristic (assuming no pruning). You do not need to show all the calculations.

| Major | CGPA | Project | Shortlisted |
|---|---|---|---|
| A | H | Y | Y |
| A | L | N | N |
| A | L | Y | N |
| A | L | N | N |
| C | H | Y | Y |
| C | L | Y | Y |
| C | H | N | Y |
| C | L | N | Y |

**Solution:**



(b) Consider the following decision tree to decide whether to go for a long drive. You wish to apply post-pruning on the tree. Draw all possible candidate pruned trees. [4]




**Solution:** 4 subtrees by pruning at the 4 internal nodes.

(c) Given a tree $T$ and a candidate pruned tree $P_i$, how will you decide whether to use $T$ or $P_i$? [2]

**Solution:** Check the accuracy on the validation set. Use $T$ or $P$ depending on which has higher accuracy.

(d) Does the pruning of a decision tree reduce variance? Justify your answer. [2]

**Solution:** Pruning reduces variance as we get a less complex model.

(e) Consider the following scenario. You have 1000 samples in the training data of a 2-class classification problem. Each sample is described by 20 features. Suppose that you use information gain as a heuristic to grow a decision tree $T$ from the training set; and you use post-pruning to prune the tree to obtain a pruned tree $P$. [2]

What can you say about the error on the *training set* of the pruned tree $P$ compared to that of $T$?

**Solution:** The error on the *training set* is likely to be higher for the pruned tree $P$ compared to that of $T$. Otherwise it is unlikely that the tree could have been expanded.

3. Consider the below dataset in Figure 1. The objective is to learn a line that best separate these classes. For this problem, assume that we are training an SVM with a quadratic kernel– that is, the kernel function is a polynomial kernel of degree 2.

   Please answer the following questions qualitatively. Give a one sentence justification for each and draw your solution in the appropriate part of the Figure at the end of the problem. Assume we are using slack penalty $C$. **Recall** the SVM optimization with slack variables is

$$\min_{\theta} \frac{1}{2} \sum_{j=1}^{d} \theta_j^2 + C \sum_i \psi_i,$$

   subject to $y_i(\theta^\top x_i - 1) \geq 1 - \psi_i \forall i$.

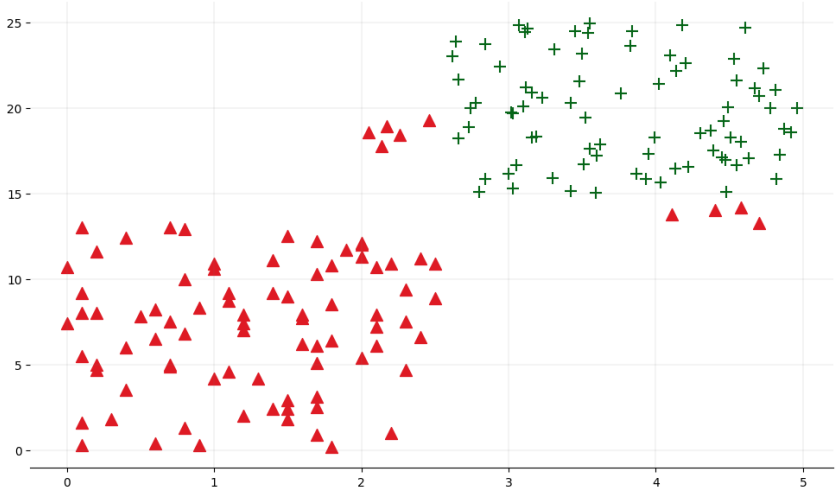



Figure 1: Dataset for SVM: + and ▲ marks two different classes.

(a) Where would the decision boundary be for very large values of $C$ (i.e., $C \to \infty$)? (remember that we are using an SVM with a quadratic kernel.) Draw on the figure and justify your answer. [3]

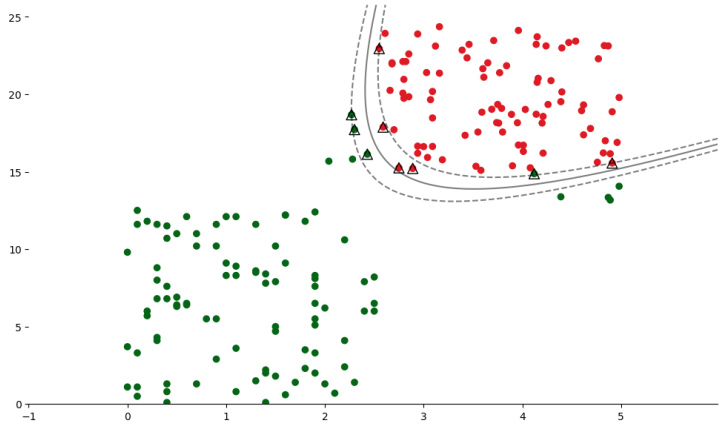



Figure 2: For Q3(a), draw the decision boundary and justify in short.

Answer: With high value of $C$, the slack variables need to be very small for minimal solution. So, the decision boundary will try to overfit to all boundary points as much as possible.

(b) Draw a data point which will significantly change the decision boundary learned for very large values of C. Justify your answer. [3]

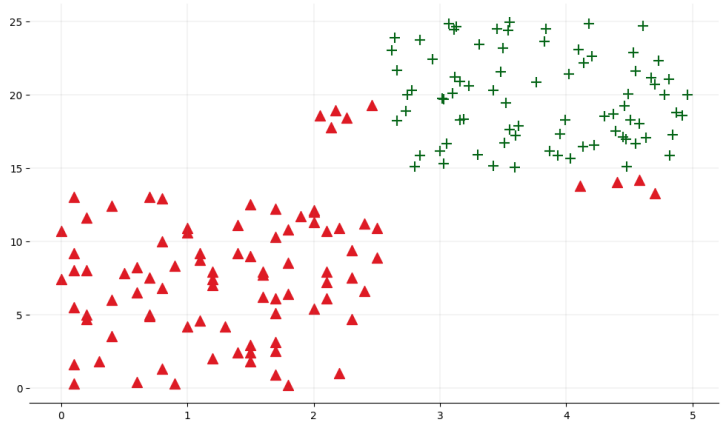



Figure 3: For Q3(b), mark the data point for large $C$ and justify in short.

Ans: A red point at (3,16) or a green point at (2.5, 13.5). With very high C, decision boundary will change to overfit. Many answers could be here. As long as students understand the effect of high $C$ value, mention a point and justify, full marks should be given.

(c) For very small $C \approx 0$, where would be the decision boundary. For this dataset, which $C$ (very large or very small) will be preferable? [4]

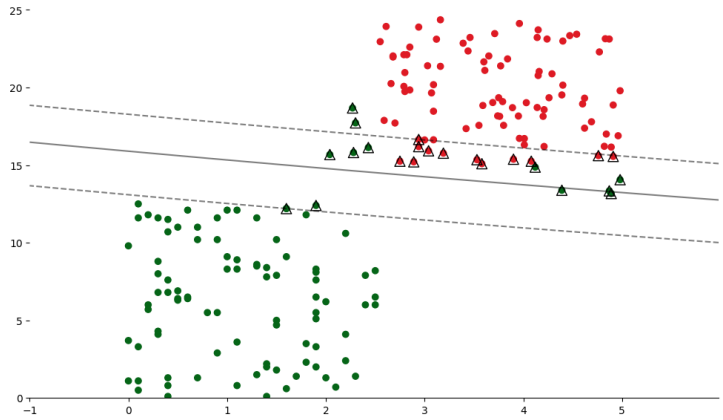



Figure 4: For (c), mark the decision boundary and justify in short.

Ans: With low $C$ value, high chance that slack variables can be high. Therefore, it will be able to ignore the outlier red dots.

(d) Suppose there are two Kernel functions $K_1(\boldsymbol{x}, \boldsymbol{x}')$ and $K_2(\boldsymbol{x}, \boldsymbol{x}')$. Show that the function $K_1(\boldsymbol{x}, \boldsymbol{x}') \times K_2(\boldsymbol{x}, \boldsymbol{x}')$ is also a kernel. To show that a function is a Kernel function, it should be factorizable into $\Phi(\boldsymbol{x})$ and $\Phi(\boldsymbol{x}')$ for some arbitrary projection function $\Phi(.)$. (*Hint*: Use the projection functions $\Phi_1(.)$ and $\Phi_2(.)$ for kernels $K_1(.,.)$ and $K_2(.,.)$. Try to derive the new projection function in terms of these. ) [3]
Ans: Assume $\phi_1$ and $\phi_2$ are two feature representation of the two kernels such that $K_1(\boldsymbol{x}, \boldsymbol{x}') = \phi_1(x)^\top \phi_1(x')$ and $K_2(\boldsymbol{x}, \boldsymbol{x}') = \phi_2(x)^\top \phi_2(x')$. The feature representation for the new kernel will be is $\phi_1(x)_i \phi_2(x)_j$, for all $i, j$.

4. (a) In linear PCA, the covariance matrix of the data $\Sigma = \boldsymbol{X}^\top \boldsymbol{X}$, is decomposed into weighted sum of its eigenvalues ($\lambda$) and eigenvectors $\boldsymbol{p}$, [4]

$$\Sigma = \sum_i \lambda_i \boldsymbol{p}_i^\top \boldsymbol{p}_i, \tag{1}$$

What is the variance of the data projected along the first dimension in terms of $\boldsymbol{p}_1$ and $\Sigma$? From this, prove that the first eigenvalue $\lambda_1$ is identical to the variance of the data projected along the first principal component.

Ans: The variance in first PC is $v = \boldsymbol{p}_1^\top \Sigma \boldsymbol{p}_1$. Since $\lambda_1$ is the eigenvector $\lambda_1 \boldsymbol{p}_1 = \Sigma \boldsymbol{p}_1 \implies \lambda_1 \boldsymbol{p}_1^\top \boldsymbol{p}_1 = \boldsymbol{p}_1^\top \Sigma \boldsymbol{p}_1 \implies \lambda_1 \times 1 = \boldsymbol{p}_1^\top \Sigma \boldsymbol{p}_1$, as $\boldsymbol{p}_1^\top \boldsymbol{p}_1 = 1$.

(b) In Figure **??** (below), there is a dataset plotted with two classes. Draw the first and second principal component directions (to your best guess). [3]

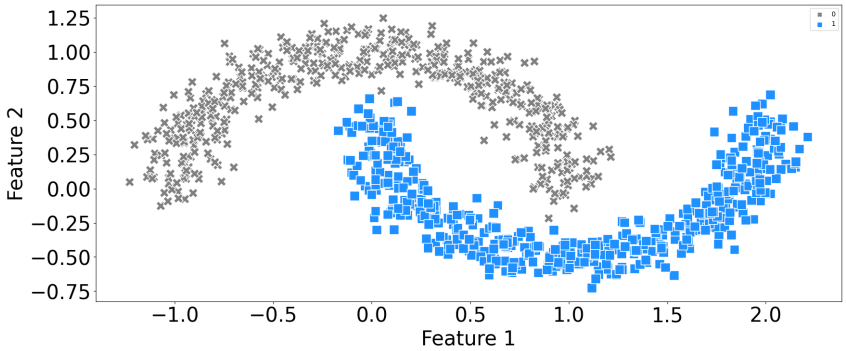



Figure 5: Dataset for PCA. For 4(c) draw the principal components and justify.

Ans: The first component of PCA will be along the x-axis (Feature 1) as it has the maximum variance along this axis. Similarly, the second component of PCA will be prependicular to the first component and hence, it will be along y-axis (Feature 2).

(c) In Figure **??**, each of the half-circles belong to a separate class. Imagine there are two classifiers learnt. [4]

- 1) Here we use PCA and take the first principal component to derive the projected data (1-D data), we further use logistic regression to classify,
- 2) We use Linear Discriminant Analysis to find the projection weight matrix $\boldsymbol{w}$, which projects from 2-D to 1 dimension. Now, we take the projected vectors and use logistic regression classifier.

Which one among the classifiers will have a better training accuracy? Explain. The classifier that has a better training accuracy, will it also have a better generalization accuracy? Explain (in short).

LDA will possibly give better accuracy than PCA because: 1. PCA will take first component along x-axis (feature 1) and in this process, it will overlap many datapoints while reducing the dimension. 2. LDA on other hand take into account the class labels and focuses on maximizing the class separability.

No, it may not translate to generalization accuracy, because the trainer may overfit the given data.

(d) Suppose that X is a discrete random variable with the following probability mass function: $P(X = 0) = \frac{2\theta}{3}, P(X = 1) = \frac{\theta}{3}, P(X = 2) = \frac{2(1-\theta)}{3}, P(X = 3) = \frac{(1-\theta)}{3}$. The following 10 independent observations were taken from such a distribution: (3, 0, 2, 1, 3, 2, 1, 0, 2, 1). What is the maximum likelihood estimate of $\theta$? [3]

SOLUTION:
Since the sample is (3, 0, 2, 1, 3, 2, 1, 0, 2, 1), the likelihood is:

$L(\theta) = P(X=3) * P(X=0) * P(X=2) * P(X=1) * P(X=3) * P(X=2) * P(X=1) * P(X=0)P(X=2) * P(X=1)$

Substituting from the probability distribution given above, we have

$L(\theta) = \prod_{i=1}^{n} P(X_i|\theta) = (\frac{2\theta}{3})^2 * (\frac{\theta}{3})^3 * (\frac{2(1-\theta)}{3})^3 * (\frac{1-\theta}{3})^2$

Clearly, the likelihood function $L(\theta)$ is not easy to maximize.

Let us look at the log-likelihood function:

$l(\theta) = log(L(\theta)) = \sum_{i=1}^{n} logP(X_i|\theta) = 5log(\theta) + 5log(1-\theta) + C$, where $C$ is a constant that represents all the terms that don't depend on $\theta$

It can be seen that the log-likelihood function is easier to maximize compared to the likelihood function.

Let the derivative of $l(\theta)$ with respect to $\theta$ be zero:

$\frac{dl(\theta)}{d\theta} = \frac{5}{\theta} - \frac{5}{1-\theta} = 0$

and the solution gives us the MLE, which is $\hat{\theta} = 0.5$