CS60050

Practice Problems

Linear Regression

Q1. Consider the following training data. Perform linear regression on this data.

x1	x2	y
0	0	0
0	1	3
1	0	4
1	1	5

a) Show the steps to find the closed form solution of the form $w_0 + w_1 x_1 + w_2 x_2 = 0$

b) Start with $w_0 = 0, w_1 = 1, w_2 = -1$. Perform two steps of iterative gradient descent and show how the parameters change. Take $\alpha = 0.5$

Logistic Regression

Q2. Consider a binary classification problem whose features are in R^2. Suppose the predictor learned by logistic regression is

$\sigma(w_0 + w_1 x_1 + w_2 x_2)$ where $\mathrm{w}_0 = 4, w_1 = -1, w_2 = 0.$

Find and plot curve along which P(class 1) = 1/2 and the curve along which P(class 1) = 0.95.

Logistic Regression

Q3. Consider a 3-class classification problem. You have trained a predictor whose input is $x \in R^2$ and whose output is $\text{softmax}(x_1 + x_2 - 1, 2x_1 + 3, x_2)$. Find and sketch the three regions in R^2 that gets classified as class 1, 2, and 3.

Q4. You are training a logistic regression model and you notice that it does not perform well on test data.

a) Could the poor performance be due to underfitting? Explain.

b) Could the poor performance be due to overfitting? Explain

Decision Tree Learning

Q5. The following table gives a data set for deciding whether to play or cancel a ball game, depending on the weather conditions.

- At the root node for a decision tree in this domain, what are the information gains associated with the Outlook and Humidity attributes?
- Draw the complete (unpruned) decision tree, showing the information gain at each non-leaf
- node, and class predictions at the leaves.

Outlook	Temp (F)	Humidity (%)	Windy?	Class
sunny	75	70	true	Play
sunny	80	90	true	Don't Play
sunny	85	85	false	Don't Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
overcast	72	90	true	Play
overcast	83	78	false	Play
overcast	64	65	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don't Play
rain	65	70	true	Don't Play
rain	75	80	false	Play
rain	68	80	false	Play
rain	70	96	false	Play

Q6. Table 1 describes positive and negative instances of people who were and were not granted credit card. Each row indicates the values observed, and how many times that set of values was observed. For example, (F, Low, +) was observed 10 times, while (F, Low, +) was observed 80 times

- Compute the sample entropy H for this training data (with logarithms base 2)?
- Calculate the information gains (IG)
  - IG(Gender) = H(Approved) − H(Approved|Gender)
  - IG(Income) = H(Approved) − H(Approved|Income)

Table 1: Credit Card Application With Two Attributes

Gender	Income	Approved	Counts
F	Low	+	10
F	High	+	95
M	Low	+	5
M	High	+	90
F	Low	-	80
F	High	-	20
M	Low	-	120
M	High	-	30

Q7. Table 2 describes the positive and negative instances of people who were and were not granted credit card. Each applicant either gets accepted (+) or rejected (-). Here each instances have three attributes: gender (F, M), income (Low, High), and age.

- Note the continuous attribute age. To increase the complexity of a decision tree by the same amount for any decision, only binary splits of the form age < A vs. age ≥ A are allowed,

Table 1: Credit Card Application With Two Attributes

Gender	Income	Approved	Counts
F	Low	+	10
F	High	+	95
M	Low	+	5
M	High	+	90
F	Low	-	80
F	High	-	20
M	Low	-	120
M	High	-	30

a) How many possible values of A do we need to consider to determine the optimal root split for the attribute age (note that some age values are repeated more than once)?

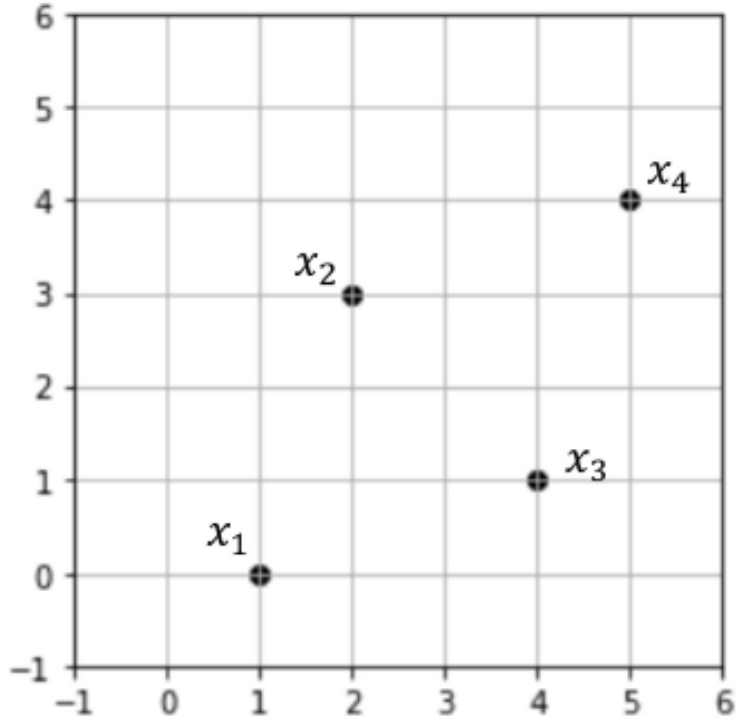
b) Draw the decision tree that would be learned by the information-gain based algorithm and annotate each non-leaf node in the tree with the information gain attained by the respective split.

c) Change one input attribute of one example in the above data set, so that the learned tree will contain at least one additional node.

d) We call a training example consistent with a decision tree if it is classied correctly be the tree. Is it possible to add new examples to the original training set which are consistent with the tree learned in (2), but nevertheless result in the ID3 algorithm run on the enlarged training set to learn a tree whose root node is different from the original trees and whose number of nodes is larger than the original trees? Justify your answer by explaining informally why this is impossible, or explaining the new data you would add.

Feature Extraction

Q8. Consider the four sample points: $\{x_1 = (1,0), x_2 = (2,3), x_3 = (4,1), x_4 = (5,4)\}$. You find that there are two eigenvectors of the covariance matrix : $(\frac{1}{\sqrt{(2)}}, \frac{1}{\sqrt{(2)}}$ with eigenvalue 16 and $(\frac{1}{\sqrt{(2)}}, -\frac{1}{\sqrt{(2)}}$ with eigenvalue 4. Use PCA to obtain a 1-dimensional projection of the data to a single dimension z. Show the 1-d axis in the figure below which contains the sample instances, and show the (approximate) projection of each sample in the figure. On the right side write down the projected points in this 1-d representation. Also write the exact coordinates of two of the reconstructed points : $\tilde{x_1}$ and $\tilde{x_2}$.



$z_1 =$

$z_2 =$

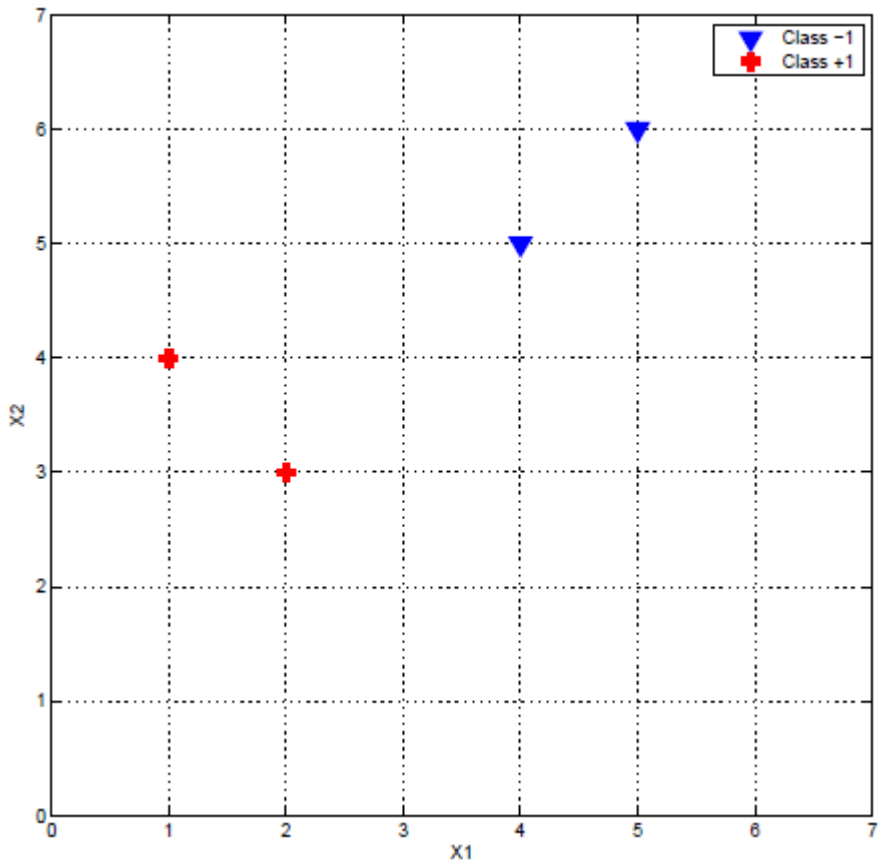
$z_3 =$

$z_4 =$

$\widetilde{x_1} =$

$\tilde{x}_2 =$

SVM



- You are training SVM (Hard Margin) on the given dataset with 4 points shown in Figure. This dataset consists of two examples with class label -1 (denoted with plus), and two examples with class label +1 (denoted with triangles).
- What is the equation of the decision surface?
- Circle the support vectors.

Properties of Kernels

Symmetry and positive semi-definiteness

A function $K: X \times X \rightarrow R$ is a valid kernel if and only if it is symmetric and positive semi-definite, that is,

1. $K(x, x') = K(x', x)$
2. Any of the following equivalent statements holds:
   i. The kernel matrix K computed on data $\mathrm{D} \subset X$ is positive definite for all D, that is, $v^T K v \geq 0 \; \forall v \in R^n$

   ii. $\sum_i \sum_j c_i c_j K(x_i, x_j) \geq 0 \quad \forall D \subset X, c_i, c_j \in R$

Kernel composition rules

Sum Rule

If $k1$ and $k2$ are valid kernels on X, then $k1 + k2$ is a valid kernel on X.

Scaling rule

- If $\lambda > 0$ and k is a valid kernel on X, then λk is a valid kernel on X.

Product rule

If $k1$ and $k2$ are valid kernels on X, then $k1k2$ is a valid kernel on X.

If $k1$ is a valid kernels on X_1, and $k2$ is a valid kernels on X_2, then $k1k2$ is a valid kernel on $X_1 \times X_2$.

Proving kernel validity

1. Proving that a kernel is valid:

1.1 Prove symmetry and positive definiteness

1.2 Find an explicit feature map $\phi(x)$, such that $k(x, x') = \phi(x)^T \phi(x')$

1.3 Derive the kernel from other valid ones using the composition rules

2. Proving that a kernel is invalid:

2.1 Find a counterexample against symmetry

2.2 Find a counterexample against positive definiteness

Sum rule

- Suppose k1 and k2 are kernels.. Then the following are also valid kernels:

$$k(x,z) = \alpha k1(x,z) + \beta k2(x,z) \text{ for } \alpha, \beta \geq 0$$

$$k(x,z) = \left[\sqrt{\alpha}\phi_1(x), \sqrt{\beta}\phi_2(x)\right]^T \left[\sqrt{\alpha}\phi_1(z), \sqrt{\beta}\phi_2(z)\right]$$

Proof of closeness of kernel functions under product

If $k1(x,y), k2(x,y)$ are both valid kernel functions, then their product $k_p(x,y) = k1(x,y)k2(x,y)$ is also a valid kernel function.

- Since $k1, k2$ are valid kernels, they must admit an inner product representation. Let Φ_1 denote the feature vector of $k1$ and Φ_2 denote the same for $k2$

$$k1(x,y) = \Phi_1(x)^T \Phi_1(y)$$
$$k2(x,y) = \Phi_2(x)^T \Phi_2(y)$$

$$k1(x,y)k2(x,y) = \left(\sum_i \phi_{1,i}(x)\phi_{1,i}(y)\right) * \left(\sum_j \phi_{2,j}(x)\phi_{2,j}(y)\right)$$

$$k1(x,y)k2(x,y) = \sum_{i,j} \phi_{1,i}(x)\phi_{2,j}(x)\phi_{1,i}(y)\phi_{2,j}(y)$$

We can define $\phi_p(z) = \phi_{1,i}(z)\phi_{2,j}(z)$ because each ϕ function outputs a scalar. Thus, we can finally write:

$$k1(x,y)k2(x,y) = \sum_k \phi_p(x)\phi_p(y)$$

Thus, the product of two kernels is a kernel.

After multiplying, our remapped feature space is effectively larger by a square.

We want to show that if the feature vectors x, y are of dimension 2, then

$$k(x, y) = x_1^2 y_1^2 + x_2^2 y_2^2$$

is a valid kernel. Prove by finding the corresponding $\phi(\ )$.

Ans. $\phi(\ x) = [x_1^2 \quad x_2^2]^T$, $\phi(\ y) = [y_1^2 \quad y_2^2]^T$