Week 10 Lab. DS-GA 1003 Machine Learning
Spring 2019
Sreyas Mohan, CDS

Some of the questions are taken from concept check questions and various other sources.

1 Multiclass

1. We are given the dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in R^2$ and $y_i \in \{1, 2, 3\}$. Using a one-vs-all methodology we have fit the corresponding score functions $f_{w_i}(x) = w_i^T x$ for $i = 1, 2, 3$ where

$$w_1 = (2, -1), \quad w_2 = (-1, 1), \quad w_3 = (-2, -2).$$

   (a) To fit each w_i we used a standard soft-margin SVM. If $\mathcal{D} = \{((0,1),2), ((1,1),1), ((-2,-2),3)\}$, what dataset was given to the SVM to fit w_3?

   (b) For each of the following new datapoints x, state which class will be predicted.
   i. $x = (-3, 2)$
   ii. $x = (1, -1)$

   (c) We want $\psi : R^2 \times \{1, 2, 3\} \to R^D$, for some D, and $\tilde{w} \in R^D$ so that

$$x \mapsto_y \tilde{w}^T \psi(x, y)$$

   gives the same classification function as the one-vs-all method described above. Give explicit values for $\tilde{w}$, $\psi(x, 1)$, $\psi(x, 2)$, and $\psi(x, 3)$ for which this is the case. If needed, you may refer to the components of x by $x = (x^1, x^2)$.

2 Decision Trees

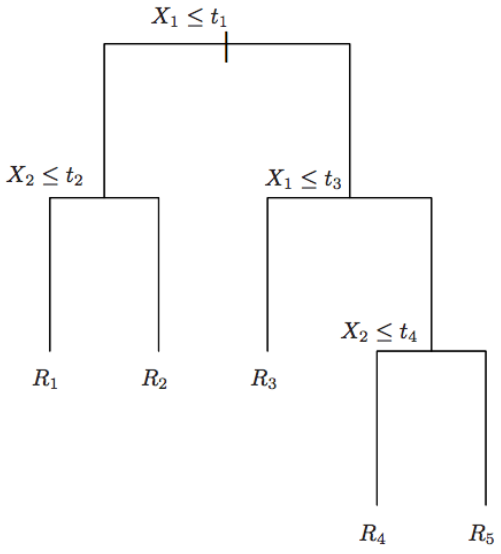

Figure 1: Tree for Question 1

2. (a) Consider a tree build on $\mathcal{R}^2$ shown in figure 1. Each leaf node represent a subset of region in $\mathcal{R}^2$ labelled in the picture as R_i. Draw a diagram showing this split.

   (b) Let $\mathcal{X}$ represent the input space of a tree and let R_i represent the region associated with the i^{th} leaf node. Which of the following statements are true?

   - ◯ $\bigcup_i R_i = \mathcal{X}$
   - ◯ $\bigcap_i R_i = \mathcal{X}$
   - ◯ $R_i \cap R_j = \phi$ for **any** $i \neq j$
   - ◯ There exits i and j such that $R_i \subset R_j$

   (c) The height of a tree is the number of edges in the longest path from the root to any leaf. What is the height of tree build on $\mathcal{R}^2$ shown in figure 1?

   (d) What is the maximum number of regions a (binary) tree of height k can have?

   (e) Which **ONE** of the following quantities best explains the worst-case running-time when making a prediction using a decision tree?

   - ◯ The number of leaves.
   - ◯ The number of training data points.
   - ◯ The number of features.
   - ◯ The height of the tree.

3. Classification with Trees.

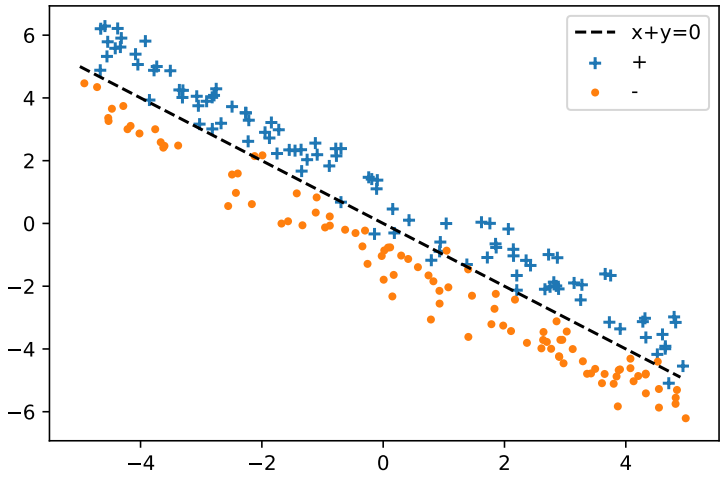

Figure 3: Dataset with two classes

   (a) (True/False) Trees assume that the data is linearly separable.

   (b) Give an upper bound on the depth needed to exactly classify n distinct points in $\mathcal{R}^d$.

   (c) Consider the data set shown in Figure 3 . How would the classification boundary look like if we build a tree on the input space $\mathcal{R}^2$?

   (d) What can you do to classify the data set with a tree of depth 2?

4. Let $\mathcal{D} = \{(1, 2), (-2, 1), (3, 2)\}$ where $\mathcal{X} = \mathcal{Y} = R$. Give an expression for a prediction function $f : R \to R$ that minimizes the square loss over the space of regression stumps (i.e. regression trees of height 1). If you prefer, you may just draw the tree, so long as it contains the same information.

5. Recall that:

1. Misclassification error: $1 - \hat{p}_{mk(m)}$
2. Gini index: $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
3. Entropy: $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$

Suppose we are looking at a fixed node of a classification tree, and the class labels are, sorted by the first feature values,

$$4, 1, 0, 0, 1, 0, 2, 3, 3.$$

We are currently testing splitting the node into a left node containing $4, 1, 0, 0, 1, 0$ and a right node containing $2, 3, 3$. For each of the following impurity measures, give the value for the left and right parts, along with the total score for the split.

1. Misclassification error.
2. Gini index.
3. Entropy.

6. **Decision Trees**:
In this question, we will build a binary classificaton tree by hand, using entropy as our impurity measure. As a reminder:

$$\text{Entropy:} \qquad H(X) = -\sum_{i=1}^{n} P(x_i) log P(x_i), \;\; X \in \{x_1, \cdots, x_n\}$$

Consider the following training data (with $(x1, x_2) \in R^2$, $\mathcal{Y} \in \{0, 1\}$).

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 6 & 6 \\ 9 & 5 \\ 1 & 2 \\ 7 & 3 \\ 5 & 4 \end{bmatrix} \qquad y = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

(a) What is the entropy at the root of the tree?

(b) Select all of the optimal split rules for the root node from the choices below.
◯ $x_1 \geq 5.5$ ◯ $x_1 \geq 7$ ◯ $x_2 \geq 4$ ◯ $x_1 \geq 7.5$ ◯ $x_2 \geq 4.5$

(c) Draw the full decision tree. You do not need to justify splits, but you do need to:

- Give the splitting rule for each tree node.
- Give probability estimates $\hat{p}$ in each leaf node.

An example is shown in figure 4:

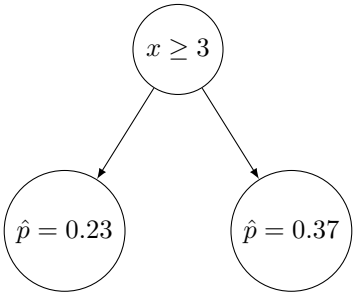

Figure 4: Example showing how to draw tree

(d) Indicate which of the following would be expected to reduce overfitting in a decision tree. Answer in general – not about the specific tree you just constructed.

- ◯ Pruning the decision tree.
- ◯ Increasing maximum tree depth.
- ◯ Decreasing the minimum impurity decrease required to split a tree node.
- ◯ Decreasing the maximum number of leaf nodes.
- ◯ Increasing the minimum number of training instances required in a leaf node.