Jack Yeung
3/31/24
B455

Assignment06

Question 1

Suppose we have a following function

$$E(w) = w^2 4w + 4$$

we want to minimize this function using gradient descent. Our goal is to find the value of w that minimizes this function. Due to time constraint we are going to run only 3 iterations for gradient descent. So, your task is to find the value w after 3rd iteration i.e. w_3 Below are the given things

- Initial value $w_0 = 0$
- Learning rate $\eta = 0.1$

Show detailed calculations for each iteration (Iteration 1, 2, 3)
Derivative of the function:

$$E'(w) = 2w - 4$$

- Iteration 1:

$$w_1 = w_0 - \eta \cdot E'(w_0) = 0 - 0.1 \cdot (-4) = 0.4$$

- Iteration 2:

$$w_2 = w_1 - \eta \cdot E'(w_1) = 0.4 - 0.1 \cdot (-3.2) = 0.72$$

- Iteration 3:

$$w_3 = w_2 - \eta \cdot E'(w_2) = 0.72 - 0.1 \cdot (-2.56) = 0.976$$

Question 2

Assume that you are working on the cancer detection problem. You have 2 classes in your problem C1 = Cancer Positive and C2 = Cancer Negative. The data point belongs to class C1 if value of discriminant function ¿ 0 or else the data point belongs to C2. Our input data points are represented by 2D data as

$$(x_i, y_i)$$

. We want to find the linear discriminant that separates these two classes. Let's assume the weight vector $w(w_1, w_2)$ is (2, -3) and threshold value w_0 is 5. For

the given points (x,y), calculate following a) Value of the discriminant function g(x, y) for 2 points $(x_1, y_1) = (1, 2) and (x_2, y_2) = (3, 4)$ b) Classify above both points into class C1 and C2 based on value of g(x,y) Show detailed calculations

Given:

- Weight vector $\mathbf{w} = (2, -3)$
- Threshold $w_0 = 5$
- Discriminant function $g(x, y) = w_1 x + w_2 y + w_0$

1. For point $(x_1, y_1) = (1, 2)$:

$$g(1, 2) = 2 \cdot 1 - 3 \cdot 2 + 5 = 1$$

Cancer Positive because discriminant is larger than 0

2. For point $(x_2, y_2) = (3, 4)$:

$$g(3, 4) = 2 \cdot 3 - 3 \cdot 4 + 5 = -1$$

Cancer Negative because discriminant is less than 0

Question 3

a) What is gradient descent and why is it used in optimization

- Gradient Descent is used when there is no analytical solution(when you cannot set the derivative to 0 and solve). Therefore we must use an iterative process, this process is done by selecting a point, moving in the opposite direction of the gradient of error. When repeating until convergence, this algorithm will find the nearest minima.

b) What is the role of the learning rate in gradient descent?

- The size of the learning rate i.e. the step is important. If the step is too large the algorithm may overstep the minima, which as the book states will "cause oscillations and even divergence". On the other hand if the step is too small, convergence will be computational expensive and will take too long.

Question 4

a) How does the linear discriminant model attribute importance to input features?

- Because the linear discriminant function is multiplied by the weight vector, we are adding priority to different dimensions/features. An easy example is of Question 2 where the weight vector was (2,-3) meaning that feature 1 has a positive correlation and feature 2 a negative one.

b) Explain the concept of ranking in machine learning and how it differs from classification and regression tasks

- Ranking in machine learning is somewhat of a combination of regression and classification. Instead of finding the discriminant or the regression formula, we must find a score value. As the book states, within an algorithm that ranks a users movies, there is a lot of difficulty using ranking algorithms as taste is nuanced and is difficult to capture with binary options such as "enjoyed" or "not-enjoyed". Ranking is also different from the two main goals of machine learning as it is not about accurately predicting the results, but more about predicting the relative relationship between instances.

Question 5

a) Explain the concept of pruning in decision trees. Discuss why pruning is important, and the methods commonly used for pruning decision trees. Additionally, outline any potential drawbacks or challenges associated with pruning decision trees, and provide insights into when pruning might not be advisable.

- pruning is vital to avoid over fitting with decision tree models. pruning can be done preemptively or posterior. in prepruning, the percentage of data points has a maximum given by a parameter. Post-pruning involves creating a full, pure tree , then reducing sub-trees that contribute to over fitting. This is quantified if the sub-model retains performance when converting sub-trees to leaf nodes. In general, prepruning tends to be quicker and postpruning is more accurate.

b) Is it possible to perform regression with decision tress? If yes how is it done. Explain with an example.

- Regression is possible using decision trees. A regression tree is very similar to a classification tree but instead the impurity measure is different. Impurity is now assessed using mean squared error from the estimated data, where we want to reduce variance to an acceptable point(which is given by a complexity parameter). With this tree to predict the target value of a new dataset we traverse the tree until we hit a leaf node. Assigning the value of this new point to an average of the y value from all of the training data points present at that leaf node.

c) Explain the concept of rule induction from decision trees. Discuss why rule-based representations are preferred for model interpretability.

- Rule induction is a way of retrieving explicit rules on the data. The algorithm performs a depth first search from the root to each leaf. Because of this, each split can be represented as a combination of if-then statement. This has greater interpretability than for example, the significance of weights in a multilayer-perceptron.

Question 6

a) Explain the concept of entropy in the context of decision trees.

1. entropy is a measure of impurity for a given dataset. specifically this is used withing decision trees to measure how effective a split is. Ideally when making decisions to further split the data, we would like to pick spots with entropy, such that you are splitting when there is not clear separation of classes.

b) Calculate the entropy of the dataset. The target variable for the dataset is Play Tennis

- The formula for entropy is:

$$\text{Entropy}(S) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

And given the probability of playing tennis being 9/14 and the probability of no being 5/14, we can plug in and simplify the formula to

$$\text{Entropy}(S) = -(9/14 \log_2(9/14) + 5/14 \log_2(5/14)$$

simplifying to out final answer of

$$\text{Entropy}(S) = .94)$$