Homework 2

16:198:461

1 Decision Trees for Classification

Generate synthetic data, each data point of the form $\underline{X} = (X_1, X_2, X_3, \ldots, X_{15})$ in the following way:

- For $i = 1, 2, \ldots, 15$, let $X_i = -1$ with probability $1/2$, and $X_i = 1$ with probability $1/2$.

For each data point, take Y as

$$Y = \text{sign}\left(0.9X_1 + 0.9^2X_2 + 0.9^3X_3 + 0.9^4X_4 + 0.9^5X_5 + N\left(0, \sigma^2\right)\right), \tag{1}$$

where σ^2 is the variance on a randomized noise term.

1) Write a program to generate data in accordance with the above, with $\sigma = 0.05$. Generate a training data set of size 5000, and a test data set of size 500. Plot the number of misclassifications on the training data and testing data, as a function of the minimum sample size you grow the tree to. What is the optimal sample size to grow the tree to, according to your data? Note - only consider odd sample sizes, so you can resolve terminal nodes by majority vote.

2) Repeat this a few times - is this optimal sample size consistent? Take the average optimal sample size, call it s.

3) Generate again a training set and testing set for the $\underline{X}$ data as above, but we want to consider how the error on the model changes as a function of σ. Plot, as a function of σ, the training and testing error on a decision tree grown to sample size s. Note, for each σ you test, you'll need to recalculate Y based on that σ, and use those Y values to grow the decision tree. How does the noise influence the effectiveness of the tree?

4) In parallel with Part 3 above, also plot as a function of σ the number of times irrelevant features show up in the tree (counting duplicates). How does the noise influence the likelihood of including irrelevant features?

2 Logistic Regression

Consider data as generated in the previous section.

1) Generating a training data set of size 5000 and a test data set of size 500 with $\sigma = 0.05$, grow a decision tree on this data, but instead of resolving terminal nodes by majority vote, let the output of the tree be the fraction of data points in the terminal node that correspond to class $Y = +1$, i.e., an estimate for the probability a data point belongs to class $Y = +1$. Plot the logistic error of the decision tree on the test data and the training data as a function of the minimum sample size you grow the tree to. What is the optimal sample size to grow the tree to, according to your data? How does it compare to the previous section, when trying to predict the class label itself?

2) Consider fitting a logistic model to the data, i.e., $F(\underline{x}) = \text{sigmoid}(w_0 + w_1x_1 + w_2x_2 + \ldots + w_{15}x_{15})$, taking $F(\underline{x})$ to be the probability that data point $\underline{x}$ belongs in class $Y = +1$. Plot the logistic error of the model on the test data and the training data as a function of time as you fit the model. Is overfitting an issue here?

3) For the best decision tree model and the best logistic model - which one is better for modeling the data? Which one is better at minimizing the influence of irrelevant features - and how can you measure this?

3 Support Vector Machines

Consider the following data set, each point of the form $(X_1, X_2), Y$:

$$\begin{aligned}(-1,-1),-1\\(+1,-1),+1\\(-1,+1),+1\\(+1,+1),-1\end{aligned} \tag{2}$$

For a Kernel of your choice (that is not the Gaussian Kernel) formulate and solve the Dual SVM on this data, and plot the resulting classification regions on the (X_1, X_2) plane.

4 Bonus: Convergence

Consider a data set generated in the following way: $X_1, X_2, \ldots, X_{50}$ are i.i.d. $N(0,1)$ random variables, $X_{51} = X_{50} + \text{sign}(X_{50}) * \epsilon$, and $Y = \text{sign}(X_{51})$.

1) Show that a perceptron exists on this data set for $\epsilon > 0$.

2) Plot the amount of time it takes to find such a perceptron as a function of $\epsilon \in [0, 2]$. What do you notice?

3) Is there a better model that should be applied here? Justify. How are you defining 'better'?