Statistics 315B Homework 1

Spring 2017

May 4, 2017

1

The following points could be derived from my demography.

- The anuual income is predicted by four main variables: householder status, age, marital status, and occupation. This list is in order of variable importance.
- Those who do not have their own homes and are under the age of 20 are expected to have a household annual income of around \$17,500 or less.
- Those who do not have their own homes and are older than 20 years are expected to have a household annual income of \$17,500 or more but less than \$27,500.
- Those who have their own homes and are currently married or living together are expected to have a household annual income of less than \$32,500 and less than about \$50,000.
- A person who owns a house and is not currently married or living together is estimated to have a household annual income of \$22,500 or more and \$32,500 or less.

1.a

Surrogate splits were used in the optimal tree I obtained. This can be confirmed with the summary () function for the fitted model variable. Surrogate splits divide the missing data if the primary splits fail to divide the data for it. When missing data comes in as input to the model, surrogate split takes over the work done by the primary split and processes it. For example, let's look at detail for node 1 at the top of the summary() output. Node 1 divides data by householder status, and surrogate split of this node divides data by age.

1.b

Blank

2

The following points could be derived from my demography.

- The house type is predicted by five main variables: persons in own household, household status, occupation, annual incom of household, and residence period. This list is in order of variable importance.
- Those who do not rent residence are expected to live in house.
- People who rent a residence and have more than three residents are expected to live in a house or apartment.
- People who rent a residence and have less than two residents are expected to live in an apartment or other type of home.
- My tree does not categorize two categories because data from people living in condominiums or mobile homes is very rare.

In classification tree, the n-fold(in my case, n=10) cross-validated error rate can be computed from a product of root node error and the last real error, which are shown as the output of printcp(). I got error estimation, the cross-validated error rate $0.40985 * 0.63238 = 0.25918$.

3

There are two causes of overfitting. First, the model may try to 'memorize' rather than 'learn' the learning data. This means that the complexity of the model is way too high. Second, distribution of test data and test data may be different in the beginning. This is interpreted as a side effect of some sort of baised sampling.

4

According to the universal approximation theorem, a particular model can construct a function that maps a set of learning sets precisely. Nevertheless, the reason for not selecting the prediction function in the class of all possible functions is that the bias-error variance trade-off can lead to high test errors when selecting a model with too high complexity.

5

For many applications, our objective will be more complex than simply minimizing the number of misclassifications. We can formalize such issues through the introduction of a loss function, also called a cost function, which is a single, overall measure of loss incurred in taking any of the available decisions or actions. Our goal is then to minimize the total loss incurred. A loss function can use a direct prediction function such as 0-1 loss, but other functions such as cross-entopy can be used as long as the given problem is solved properly.

6

Suppose hypothesis $h : X \to Y$. Risk under h is $R(h) = \mathbf{E}[L(h(x), y)] = \int L(h(x), y)P(x, y)dxdy$. The problem is that we do not know exact $P(x, y)$. Therefore, ERM algorithm which works only on the training set (x_i, y_i) minimizes the empirical risk $R_{emp} = \frac{1}{m}\sum_{i=1}^{m} L(h(x_i), y_i)$.

However, overfitting problems can not be avoided because the learning process is applied only to the training set. This is also directly related to the bias-variance trade-off problem. Therefore, the penalization risk that gives penalty to the complexity of the model may be given as follows. $R_{pen}(h) = R_{emp}(h) + \lambda\phi(f(h))$ where ϕ is nonnegative function.

7

Misclassification risk is the following :

$$L(y_k, \widehat{y}_k) = \sum_{k=1}^{K} I(y_k \neq \widehat{y}_k)$$

where y_k is the predicted value and $\widehat{y}_k$ is the true value of the outcome y.

Classification error rate is the follwoing :

$$\frac{1}{K}\sum_{k=1}^{K} I(y_k \neq \widehat{y}_k) = \frac{L(y_k, \widehat{y}_k)}{K}$$

Thus, misclassification risk reduces the classification error rate.

8

In Words

9

On Handwriting

10

A method of surrogate splits has been proposed to deal with missing data. The idea of a surrogate split at a given node τ is that we use a variable that best predicts the desired split as a substitute variable on which to split at node τ. If the best-splitting variable for a future observation at τ has a missing value at that split, we use a surrogate split at τ to force that observation further down the tree, assuming, of course, that the variable defining the surrogate split has complete data.

11

On Handwriting

12

..

13

..

14

Again, it is necessary to focus on the bais-variance trade-off. Enlarging F will reduce *MSE* because it reduces the restriction on $g(x)$. However, this increases the complexity of the model. Thus, bias-variance trade-off reduces the bias and eventually leads to increased variance. This means overfitting of the model. Conversely, decreasing the size of F increases the restriction to $g(x)$, thus increasing *mse* and reducing the complexity of the model. Thus, the bias-variance trade-off reduces variance and increases bais. Therefore, this change can cause underfitting of the model.

15

Linear combination splits have the advantage of increasing the predictive power. However, there is a disadvantage in that the interpretability, which is the maximum advantage of the tree-based model, is significantly degraded.

16

The advantage of the multi-way splitting strategy is that it can be fitted with a shallow depth. The disadvantage is that at the higher level, too much data can be categorized too quickly, so underfitting can occur due to insufficient data. Moreover, multiway split can be interpreted as a series of binary splits.