IE406 Final_Exam

20161190 Minjae Lee

1.

test = "sale price discount"

$P(\text{non-spam}) = \frac{2}{5}$, $P(\text{spam}) = \frac{3}{5}$ $\quad P(y|X) = \frac{P(X|y)P(y)}{P(X)}$

$P(\text{price}|\text{non-spam}) = \frac{2+1}{(2+1)+(1+1)+(0+1)} = \frac{3}{6}$

$P(\text{sale}|\text{non-spam}) = \frac{1+1}{(2+1)+(1+1)+(0+1)} = \frac{2}{6}$

$P(\text{discount}|\text{non-spam}) = \frac{0+1}{(2+1)+(1+1)+(0+1)} = \frac{1}{6}$

$P(\text{price}|\text{spam}) = \frac{0+1}{(0+1)+(3+1)+(4+1)} = \frac{1}{10}$

$P(\text{sale}|\text{spam}) = \frac{3+1}{(0+1)+(3+1)+(4+1)} = \frac{4}{10}$

$P(\text{discount}|\text{spam}) = \frac{4+1}{(0+1)+(3+1)+(4+1)} = \frac{5}{10}$

$P(\text{test}/\text{non spam}) = \frac{2}{6} \times \frac{3}{6} \times \frac{1}{6} = \frac{6}{256} = \frac{1}{36}$

$P(\text{test}/\text{spam}) = \frac{1}{10} \times \frac{4}{10} \times \frac{5}{10} = \frac{20}{1000} = \frac{1}{150}$

$P(\text{non spam}/\text{test}) = \frac{\frac{1}{36} \times \frac{2}{5}}{P(\text{test})} = \frac{1}{P(\text{test})} \times \frac{1}{90}$

$P(\text{spam}/\text{test}) = \frac{\frac{1}{150} \times \frac{3}{5}}{P(\text{test})} = \frac{1}{P(\text{test})} \times \frac{3}{250}$

$P(\text{non spam}/\text{test}) < P(\text{spam}/\text{test}) \Rightarrow$ test case classify by Spam

$0.0111/P(\text{test}) \qquad 0.0012/P(\text{test})$

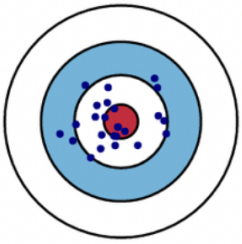
2.

1) False

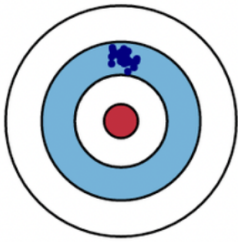
When the dataset size is small, variance is more important. Because even small changes in the training set can affect great at parameter value . So high bias low variance is better than low bias high variance classifiers.

2) True

When the dataset size is large, variance is relatively less important because there is so much data that we can accurately describe the distribution of the data. So low bias and high variance classifiers is better than high bias and low variance classifiers.



Low bias, high variance.



High bias, low variange

3.

Overfitting refers to a phenomenon that learns data models so excessively that the training data predicted well but does not predicted well with the test dataset. As explained above, the problem is that the performance is not good when testing with actual test dataset.

There are several solutions to overfitting. Overfitting occurs when M>D, that the complexity of the model's parameters is M, and the complexity of the data is D. First, there's a way to increase the number of data. This has the meaning of increasing D further. Secondly, there is a way to reduce the complexity of the model. This has the meaning of reducing M. Reducing the complexity of the model for the same data results in generalization by learning general patterns from the data. Third method, we can do the regularization to reduce the coefficient of the model. If the coefficient is large, it may react sensitively to the value and have a high error. To prevent this, the coefficient of the model is made close to zero.

4.

1) Standard random forest algorithms do not take all features into account. So this algorithm's data bias may increase, but variance is reduced through averaging. This Randomness random forests also have small variances instead of large biases. Furthermore, this randomness random forest adds more randomness in the learning process, allowing to discover a better decision tree. Randomness random forest algorithms do not need to compute best thresholds. Therefore, randomness random forest algorithm learning is relatively faster than other method.

2) The randomness random forest of the above description was uniformly selected between the minimal and maximal value for the threshold. But extreme(minimal and maximal value) thresholds

can interference with training. So, I propose the novel type of random forest algorithm that select the threshold at normal distribution. Then the extreme value will be selected relatively lower than uniformly distribution.

5.

points)

X_1	X_2	X_3	Y	
1	1	0	Positive	1/5
0	1	1	Positive	1/5
1	0	1	Positive	1/5
0	1	0	Negative	1/5
0	0	1	Negative	1/5

=> initialize same weight

x1

x_1 | y	positive	negative
1	2	0
0	1	2

x2

x_2 | y	Pos	Neg
1	2	1
0	1	1

x3

x_3 | y	pos	Neg
1	2	1
0	1	1

Let's calculate Gini index,

Case x1) $\frac{2}{5}\left(1-\left(\frac{2}{2}\right)^2-\left(\frac{0}{2}\right)^2\right)+\frac{3}{5}\left(1-\left(\frac{1}{3}\right)^2-\left(\frac{2}{3}\right)^2\right)=\frac{4}{15}$ => Smallest

Case x2) $\frac{3}{5}\left(1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\right)+\frac{2}{5}\left(1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2\right)=\frac{7}{15}$

case x3) $\frac{3}{5}\left(1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\right)+\frac{2}{5}\left(1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2\right)=\frac{7}{15}$

=> So use x1 first stump

Amount of say = $\frac{1}{2}\log\left(\frac{1-\frac{1}{5}}{\frac{1}{5}}\right)\approx 0.301$

x_1 | y	positive	negative
1	2	0
0	1	2

X_1 is 1 then y is positive

X_1 is 0 the y is negative

X_1	X_2	X_3	Y	Weight	Error	New weight	Normalize weight
1	1	0	P	1/5	no	0.148	0.112
0	1	1	P	1/5	Occur	0.27	0.313
1	0	1	P	1/5	no	0.148	0.112
0	1	0	N	1/5	no	0.148	0.112
0	0	1	N	1/5	no	0.148	0.112

New sample weight = $\frac{1}{5} \times e^{0.301} \approx 0.27$ (error occur)

$= \frac{1}{5} \times e^{-0.301} \approx 0.148$ (error no)

Iterate this method by new data set

X_1	X_2	X_3	Y	Weight
0	1	0	P	1/5
0	1	1	P	1/5
1	1	1	P	1/5
0	1	0	N	1/5
0	0	1	N	1/5

=> Calculate Gini Index

X_1 case) $\frac{4}{10}$

X_2 case) $\frac{3}{10}$ => smallest value

X_3 case) $\frac{7}{15}$

Select X_2 and total error is $\frac{1}{5}$

So, amount of say also ≈ 0.301

Same way. New sample weight => 0.27 , 0.148

↓ ↓

Normalized weight => 0.313 0.112

Iterate this method by new data set

x_1	x_2	x_3	Y	Weight
1	1	1	P	1/5
0	1	1	P	1/5
0	0	1	P	1/5
0	1	0	N	1/5
0	0	1	N	1/5

=) Same way calculate Gini index

x_1 case) $\frac{4}{10}$

x_2 case) $\frac{5}{10}$

x_3 case) $\frac{3}{10}$ => lowest

=) Select x_3 and total error is $\frac{1}{5}$.

So, amount of say also ≈ 0.301

Same way. New sample weight => 0.127 , 0.148

↓ ↓

Normalized weight => 0.313 0.102

=) So, all of the amount of say is 0.301 in my new dataset

the test case $x_1=1$, $x_2=1$, $x_3=1$ will be classify by my ada boost is

Positive

6.

They are both semi-supervised learning.

However, they have different ways of making datasets. First, self training is the most representative method of semi-supervised learning. It is also called wrapper methods. Self-training is the prediction of unlabeled data after learning the model with labeled data. It simply labels high probability values

as priority. However self-training is more sensitive than co-training to mistakes because self-training uses one model, but co-training uses multiple models.

In the case of co-training, one data is viewed from two distinct. In other words, features of the data are separated and used in different models. And these two models teach each other and train repeatedly. In other words, we can learn classifiers from each of the two independent groups, and increase the small amount of labeled data using unlabeled data to create more training sets.

7.

These data are in two-dimensional space. So let's consider two classifiers that feature is x-axis and y-axis each classifier. Then, one is classify the unlabeled data by the value of x. The other is classify the unlabeled data by the value of y. When unlabeled data are classified with the two classifiers separately, their confident sets are produced and added into labeled data of the others. Repeat this process until the unlabeled point A has same result from each classifier. Then the point A has label black circles or hollow circles.

8.

Reinforcement learning is a slightly different concept from supervised learning and unsupervised learning. Reinforcement learning does not rely on static datasets, but rather operates in a dynamic environment and learns from the collected experiences. Reinforcement learning is learning what action is optimal to take in the current state. Each action is given a reward in the external environment, and learning proceeds in the direction of maximizing this reward.

Supervised learning is learning with data which have correct answers. When the input value (X data) is given, a label (Y data) for the input value is given to learn. This learning is approximate a function that minimizes loss.

9.

Destination path

Exploitation : go to the shortest path that provided navigation

Exploration : go to the unknown path not sure if it's short path or not.

Go restaurant

Exploitation : go to the famous(best review point) restaurant

Exploration : go to new restaurant not sure if it's good or not

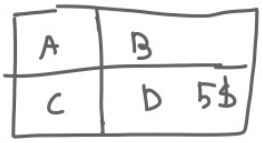
Play Game

Exploitation : play proven(best review point) game with a lot of users.

Exploration : play game that doesn't have enough evaluation because many people haven't. played it yet

10.

10_1)



	A	B	C	D
N	0	0	0	0
S	0	0	0	0
E	0	0	0	0
W	0	0	0	0

⇒ 1 iteration

	A	B	C	D
N	0	0	0	0
S	0	0	0	0
E	0→-1	0	0	0
W	0	0	0	0

⇒ 2 iteration

⇒

	A	B	C	D
N	0	0	0	0
S	0	0→4	0	0
E	0→-1	0	0	0
W	0	0	0	0

1 iter ⇒ $Q(A, East) = 0 + [-1 + 0.5 \max(0,0,0,0) - 0] = -1$

2 iter ⇒ $Q(B, South) = 0 + [4 + 0.5 \max(0,0,0,0) - 0] = 4$

10-2)

	A	B	C	D
N	0	0	0	0
S	0	4	0	0
E	-1	0	0	0
W	0	0	0	0

⇒ 1 iteration

	A	B	C	D
N	0	0	0	0
S	0	4	0	0
E	-1→1	0	0	0
W	0	0	0	0

→ 2 iteration

⇒

	A	B	C	D
N	0	0	0	0
S	0	4→4	0	0
E	1→1	0	0	0
W	0	0	0	0

1 iter ⇒ $Q(A, East) = -1 + [-1 + 0.5 \times \max(4,0,0,0) - (-1)] = 1$

2 iter ⇒ $Q(B, South) = 4 + [4 + 0.5 \times \max(0,0,0,0) - 4] = 4$

11.

1)

The two methods have pros and cons. However, in this situation, I think it is better to use content-based filtering. Since the bookstore recently launched, there is not enough rating data yet. At state (bookstore has over one million books but rating data is not enough), it occur wasting many memory because there are many parse when use collaborative filtering. Therefore, in this situation, I think content-based method which recommend a similar book if a user prefers a specific book is better

2)

Pattern Recognition and Machine Learning" is a book in the online bookstore with 500,000 ratings whose average rating is 1.0 higher than global average. -> 3.8 + 1.0 = 4.8

Ms. Kim, who is a faithful customer, has rated 500 books and her average rating is 0.5 stars higher than the average users' ratings. 4.8 + 0.5

Estimated rating: 5.3

12.

Content-based filtering and collaborative filtering are typically used for recommended systems. Both of these methods can make the user biased to one side. Content-based filtering is a method of recommending an item with content similar to that item if the user prefers a specific item. In other words, if user often watch one side of the news, only news with that tendency will be recommended. There are many types of collaborative filtering, but similar results come out. Collaborative filtering is a method of recommending a good evaluation by a user similar to me based on the user's behavior. If a user often watches A-tendency news, other users who watch A-tendency news will recommend this news to the user. As a result, the recommendation system of both methods makes the user biased to one side.