CS460200 Introduction to Machine Learning Exam 2

November 3, 2022. 10:10~11:00

1. Naive Bayes works well in most circumstances and is usually compared as a baseline. A major problem with Naive Bayes during implementation is that when the number of attributes is huge, the overall product (P(Y=C|X)) approaches zero (since the probability always falls between 0 and 1). Once the product is so tiny that it cannot be represented in double, the algorithm fails to work. Would you be able to come up with a practical solution to this critical issue? Please choose the most appropriate method to solve this issue.
   (A) To ensure the output is always equal to or larger than 0, replace multiplication with addition.
   (B) Once all the probabilities have been multiplied, normalize the obtained products between all classes to be between 0 and 1.
   (C) By adding a small epsilon (i.e., 1e-9) to all probabilities before multiplication, the products will not become zero.
   (D) By using log probabilities, we can ensure that our output will always be non-positive.
   (E) When calculating the probabilities, we only consider the numerator so that the probabilities will always be larger than 0.

   Ans: (D)
   Explanation:
   (A), (E): Violate Naive Bayes theorem
   (B), (C): The products still cannot be represented in double.

2. Ivan wants to go out and play, but he is sick. He is afraid that he got covid and will infect others when he goes out. He used to decide whether he got covid or not based on four attributes. Using a naïve Bayesian assumption, what is the possible outcome when Ivan feels sleepy, has a sore throat, and is coughing but does not have a fever? (X'=(Sleepy=Yes, Sore throat=Yes, Cough=Yes, Fever=No))

	Sleepy	Sore throat	Cough	Fever	Covid
Patient 1	No	No	Yes	Yes	Yes
Patient 2	No	Yes	No	Yes	No
Patient 3	Yes	Yes	Yes	Yes	Yes
Patient 4	Yes	No	Yes	No	No
Patient 5	No	Yes	No	Yes	Yes
Patient 6	Yes	Yes	No	No	Yes
Patient 7	No	Yes	Yes	No	Yes
Patient 8	No	No	No	No	No

   (A) Please provide the answer P(Yes|X') (3%) and P(No|X') (3%) (Estimate to the third decimal place. e.g., 0.001).

Ans: P(Yes|X') = 0.048, P(No|X') = 0.009
Explanation:
P(Covid=Yes) = ⅝, P(Covid=No) = ⅜, P(Sleepy = Yes | Covid = Yes) = ⅖,
P(Sleepy = No | Covid = Yes) = ⅗, P(Sleepy = Yes | Covid = No) = ⅓,
P(Sleepy = No | Covid = No) = ⅔, P(Sore throat = Yes | Covid = Yes) = ⅘,
P(Sore throat = No | Covid = Yes) = ⅕, P(Sore throat = Yes | Covid = No) = ⅓,
P(Sore throat = No | Covid = No) = ⅔, P(Cough = Yes | Covid = Yes) = ⅗,
P(Cough = No | Covid = Yes) = ⅖, P(Cough = Yes | Covid = No) = ⅓,
P(Cough = No | Covid = No) = ⅔, P(Fever = Yes | Covid = Yes) = ⅗,
P(Fever = No | Covid = Yes) = ⅖, P(Fever = Yes | Covid = No) = ⅓, P(Fever = No | Covid = No) = ⅔
P(Covid = Yes | X') = ⅝ * ⅖ * ⅘ * ⅗ * ⅖ = 6/125 = 0.048
P(Covid = No | X') = ⅜ * ⅓ * ⅓ * ⅓ * ⅔ = 1/108 ≈ 0.009

(B) Tell whether Ivan can go out or not.
Ans: Cannot.

3. Given the instances in the table below, we want to split data into the left subtree and the right subtree with the feature "glucose_apache" and the corresponding threshold "128". Please calculate the entropy before split, the entropy after split, and the information gain with this split combination [glucose_apache, 128]. (Round the number to the third decimal place. e.g., 0.001)
Note: Please calculate entropy with log function base 2.

patient_id	age	BMI	glucose_apache	diabetes_mellitus
32546	18	38.830	101	0
129292	39	22.721	93	0
80744	60	20.830	103	0
32716	66	25.396	200	1
45449	78	26.588	495	1
88856	70	25.859	210	1
111706	73	26.879	238	1
41226	43	33.967	102	0
70952	25	48.512	159	0
104968	68	35.349	99	0
120346	73	39.409	39	0
64196	60	19.531	80	1
41789	63	30.586	93	0
67372	63	46.364	153	1
37134	47	42.892	238	0

Ans: Entropy before = 0.971, Entropy after = 0.693, Information gain = 0.278
Explanations:
Entropy before

$= -P(diabetes_{mellitus} = 0) * \log_2[P(diabetesmellitus = 0)] - P(diabetes_{mellitus} = 1) * \log_2[P(diabetesmellitus = 1)]$

$= -\frac{9}{15} * \log_2 \frac{9}{15} - \frac{6}{15} * \log_2 \frac{6}{15} = 0.4421793 + 0.5287712 = 0.9709505 \approx 0.971$

Entropy after
= [(-0.125)*(-3) - 0.875*(-0.1926)] * 8/15+ [(-0.71428571) * (-0.485426) - 0.285714*(-1.807356)] *7/15
= (0.375 + 0.168525) * 8/15 + (0.34673286 + 0.51638691) * 7/15
= 0.28988 + 0.40278914 = 0.69266914 ≈ 0.693

Information gain = 0.971 – 0.692 = 0.278

4. In the table below, the second column represents the real label, and the third column is the result predicted by the model. Please calculate the *Precision*, *Recall*, and *F1-Score* to evaluate the model's performance. (Round the number to the third decimal place. e.g., 0.001)

patient_id	"*diabetes_mellitus*" (real label)	prediction (predicted label)
25312	1	0
59342	0	0
50777	0	1
46918	1	1
34377	1	1
74489	0	0
49526	1	1
50129	0	1
10577	0	0

Ans: Precision = 0.6, Recall = 0.75, F1-Score = 0.667
Explanations:
TP = 3
TN = 3
FP = 2
FN = 1
Precision = TP / (TP + FP) = 3 / (3+2) = 3/5 = 0.6
Recall = TP / (TP + FN) = 3 / (3+1) = 3/4 = 0.75
F1-Score = 2 * (Precision*Recall) / (Precision+Recall)
= 2 * (3/5 * 3/4) / (3/5+3/4)
= 0.6666666… ≈ 0.667

5. Which of the following description(s) is(are) true?
   (A) Compared to the decision tree, the random forest is less prone to overfitting.
   (B) Reducing the max_depth is a solution to deal with overfitting while training a decision tree.
   (C) Decision Tree Classifier can only perform binary classification on a dataset.
   (D) The Decision Tree algorithm is an unsupervised learning algorithm.
   (E) Decision Tree Algorithm can be used for both Regression and Classification problems.

   Ans: (A)(B)(E)
   Explanation:
   (C) can also perform multi-classes classification
   (D) it's a supervised learning algorithm

6. In assignment 2, how can we properly reduce some redundant nodes when building the decision tree without changing the prediction results from the model before trimming ?
   (A) Remove the limit on the maximum depth of the decision tree.
   (B) Compare the left subtree and the right subtree before constructing two branches.
   (C) Use other criteria (e.g., Gini impurity) to measure the quality of a split.
   (D) Select less features to build a decision tree.
   (E) Increase the value of "min_samples_split" while training a decision tree.

   Ans: (B)

7. Which of the following description(s) about Support Vector Machines (SVM) is(are) true?
   (A) If the training data is linearly separable, support vectors are the data points that lie closest to the decision surface.
   (B) We can't use SVM as the training method if there is no hyperplane that can separate data in the input space.
   (C) We can apply kernel functions that map the training data to a lower-dimensional space where the training data is separable.
   (D) The Hard Margin Classification tends to cause the overfitting problem because it tries to classify all training data correctly.
   (E) When applying SVM, we don't need to know the mapping between the input and feature spaces

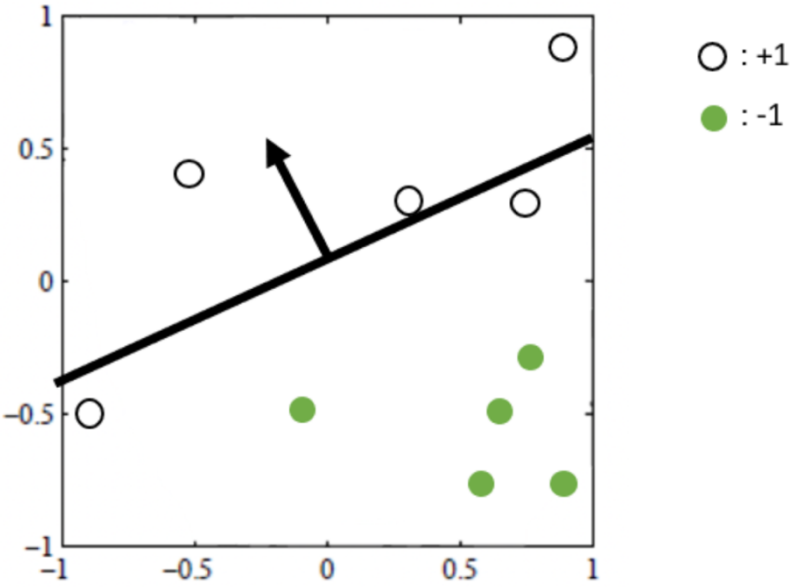
   Ans: (A)(D)(E)
   Explanations:
   (B)(C) We can use kernel functions to map the training data to the higher-dimensional space where the training data is separable

8. Which of the following statement(s) is(are) correct ?
   (A) Bagging trains classifiers in a sequence and Boosting trains classifiers independently in parallel.
   (B) Bagging can increase the complexity of under fitting models and Boosting can decrease the complexity of overfitting models.
   (C) Boosting algorithm focus on data cases which are correctly classified and increase the weight of them at each round to get a classifier.
   (D) For Boosting, we weigh the classifiers we get during the boosting process and combine them into a powerful classifier.
   (E) XGboost can use different tricks to make learning more efficient, such as regularization, pruning the tree backwards or cross-validation.

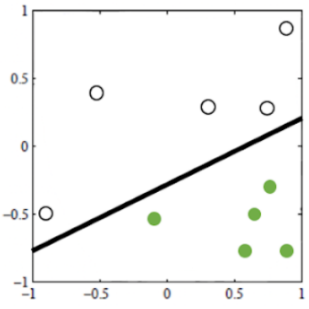
   Answer: (D)(E)
   Explanation:

(A) Bagging trains classifiers independently in parallel and Boosting trains classifiers in a sequence.
(B) Bagging can decrease the complexity of models and Boosting can increase the complexity of models.
(C) Boosting algorithm focus on data cases which are wrongly classified and increase the weight of them at each round to get a classifier.

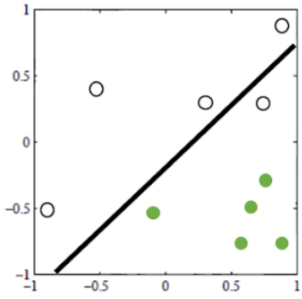
9. What is(are) the possible plot(s) for the next step using the perceptron algorithm?



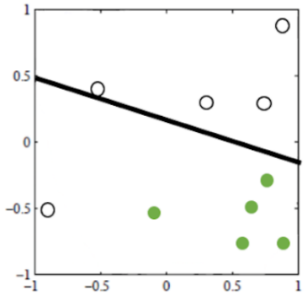
(A)



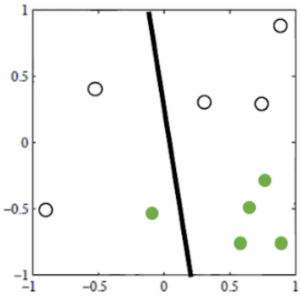
(B)



(C)

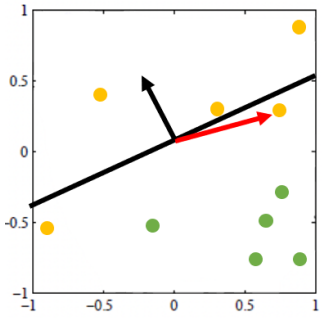


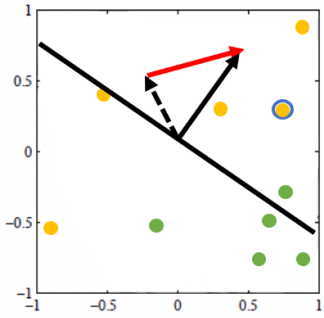
(D)



Ans: (C)(D)
Explanation:

Answer 1:





Answer 2: