CITY UNIVERSITY OF HONG KONG

Course code & title : CS5489 Machine Learning: Algorithms & Applications

Session : Midterm, Semester B 2023

Time allowed : Two hours (Mar 7th, 7:00pm-9:00pm)

This question paper has 10 pages (including this cover page).

1. This paper consists of 13 questions.
2. Answer ALL questions.
3. Write your answers in this question paper.

*This is a **closed-book** examination.*

Candidates are allowed to use the following materials/aids:

One A4 page (single-sided only) of handwritten notes with physical pen or pencil. Digital version or print of digitial verion is not allowed.

Materials/aids other than those stated above are not permitted. Candidates will be subject to disciplinary action if any unauthorized materials or aids are found on them.

Student EID: ______SOLUTIONS__________________

Student ID: ________________________________

Seat Number: ________________________________

Multiple Choice/Selection Questions (30 points)
5 marks each. *For a multiple selection question, an incorrect answer will be penalized 5/K marks, where K is the number of correct answers. If more incorrect answers are given than correct answers, the marks will be 0.*

Q1. Which statements about generative classification model are correct? (select all that apply)
A) They estimate probability distributions of features from each class.
B) It is hard to add prior knowledge to the classifier.
C) It can only work with binary classification problems.
D) Selecting different probability distributions can obtain different classification performance.
E) They predict the class with the largest class conditional densities.

Answer: <A, D>

Q2. Which statements are correct about the logistic regression? (select all that apply)
A) Logistic regression has a closed-form solution based on the MLE formulation, and thus it can be efficiently solved.
B) To prevent overfitting, we can add a prior distribution on its parameters using a Gaussian distribution.
C) Logistic regression is used for modelling the class posterior probability, which is different from generative classifiers that model the class-conditional distribution.
D) The regularization parameter of the logistic regression function can be selected via cross-validation.
E) L2 regularization is only used to stabilize the training, and it has no effects on the weights.

Answer: <B, C, D>

Q3. Which statements about support vector machines (SVMs) are correct? (select all that apply)
A) If the hyperparameter *C* is set to infinity, all the training points will be classified correctly if they are linearly separable.
B) Assuming there are three support vectors, the boundary will not change when one of the support vectors is moved.
C) If an SVM underfits the training data, then mapping the data to a lower-dimensional feature space will improve the classification performance.
D) The basic form of SVM is a convex quadratic programming problem.
E) The training data can always be correctly classified with a kernel SVM with RBF kernel function.

Answer: <A, D, E>

Q4. Which of the following are valid positive semi-definite kernel functions (select all that apply):

A) k(**x**,**z**)=$\frac{\boldsymbol{x}^T\boldsymbol{z}}{||\boldsymbol{x}||||\boldsymbol{z}||}$

B) k(**x**,**y**)=$\int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ where $p(\boldsymbol{x}|\boldsymbol{z}), p(\boldsymbol{y}|\boldsymbol{z})$ are conditional distribution, $p(\boldsymbol{z})$ is marginal distribution.

C) C, k(**x**,**y**)=$\begin{cases} 1, & -1 < \boldsymbol{x}^T\boldsymbol{y} < 1 \\ 0, & otherwise \end{cases}$

D) k(**x**,**y**)=sin($\boldsymbol{x}^T\boldsymbol{y}$)

E) k(**x**,**y**)=tanh($\boldsymbol{x}^T\boldsymbol{y}$)

Answer: <A, B>

Q5. Which statements are true about ensemble methods? (select all that apply)

A) Using fewer decision trees with higher depth is better than using more “stumps” for weak learners in Adaboost.

B) Using linear classifiers like logistic regression for bagging is a good choice because they have low computation cost.

C) Random Forest can be parallelized in both training and testing.

D) In the Adaboost training process, once a data point is correctly classified in the weak learner $h_t(x)$, it will be neglected by the successive weak learners h_{t+1}, h_{t+2}, …

E) For a random forest model, increasing the number of decision trees generally helps with reduce overfitting.

Answer: <C, E>

Q6. Considering linear regression using both L1 and L2 as regularization terms:

$$\min_{\boldsymbol{w},b} \alpha|\boldsymbol{w}| + \beta||\boldsymbol{w}||^2 + \sum_{i=1}^{N}\left(y_i - f(x_i)\right)^2$$

Which properties does it have (select all that apply)?

A) It treats all weights equally.

B) The penalty term focuses more on reducing large weights.

C) It has a closed-form solution.

D) It can perform feature selection and shrinkage simultaneously.

E) The elements of **w** are always positive.

Answer: <B, D>

Discussion Questions (70 marks)
10 marks each question.

Q7. You are working for a “smart pen” company, which can scan text using a handheld pen device. Your job is to build a model to classify the text the pen reads into different sentiment categories, such as “happy”, “sad”, “neutral”, which will then turn on different lights on the pen as a special effect. Because of the limited processing power on the pen, you will use a generative model classifier with bag-of-words model. A training set is provided, consisting of 5000 sentences and their class labels. Your initial classifier is Naïve Bayes Bernoulli model, but the performance is poor. How could you improve the model performance? (List 4 improvements)

[Any 4, +2.5 points each. -2.5 points if a reason is not correct]
1) We can try other more appropriate probability distributions, like multinomial.
2) Using n-gram bag-of-words
3) Perform text preprocessing (stemming, lemmatisation, and removing numbers or punctuation, etc.) before bag-of-words.
4) adjust the Laplace smoothing hyperparameter on the probability distributions.
5) adjust the vocabulary length.
6) try different representations, e.g., TF, TF-IDF.

Using word vector representation is a not-quite correct answer, but it will incur heavy processing -- we need to store the vector representations for all tokens, then we need to perform vector-vector multiplication for a linear classifier. Can give +1.5 point for it.

Q8. Describe the major similarities and differences between logistic regression and naïve Bayes Gaussian classifier with shared variance parameter. List at least 5 similarities/differences (5 total).

[+2 marks for correct item, -2 marks for wrong]

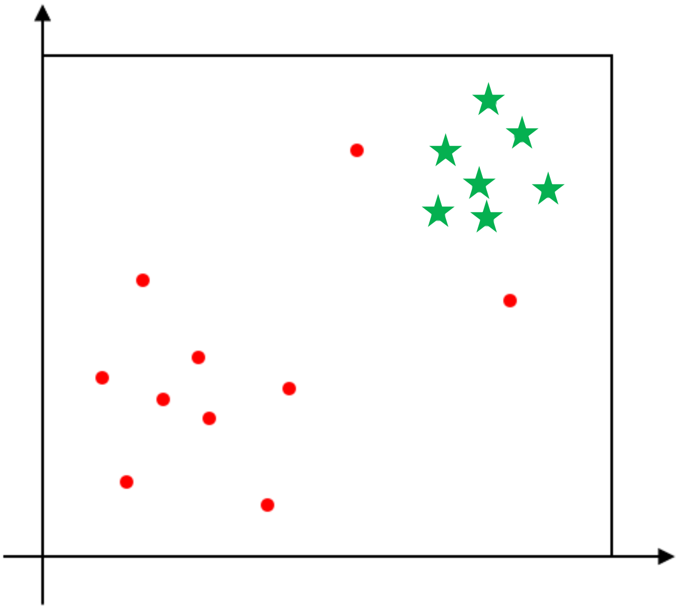
Similarities:

1) Both are linear classifiers.
2) Both use maximum likelihood estimation for learning.
3) Both can be extended to multi-class classification.
4) Both can include regularization terms in the form of priors on the weights/means.

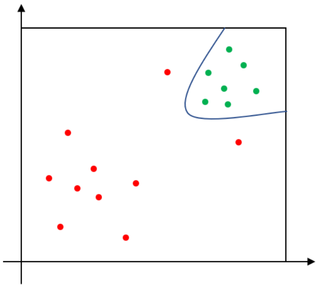
Differences:

1) Naïve bayes is a generative model, while logistic regression is a discriminative model.
2) For binary classification, naïve Bayes has 2*D+1 parameters (2 means and shared variance), while LR has D+1 parameters (w & b).
3) Naïve Bayes maximizes the class conditional likelihood, while LR maximizes the posterior probability.

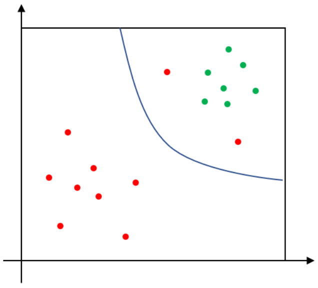
Q9. Consider the training data in the below figure, and the kernel SVM with quadratic polynomial kernel function. (a) Draw the decision boundary when $C \rightarrow \infty$? (b) Draw the decision boundary for $C \rightarrow 0$? (c) Which decision boundary would be better on the test data? For each part, give your reasons.



(a) [4 marks, 2 for answer, 2 for reason] $C \rightarrow \infty$ The penalty of misclassified sample is very large, so the model will classify all data correctly.



(b) [4 marks, 2 for answer, 2 for reason] $C \rightarrow 0$ The model tries to make a large margin, while a small number of samples are allowed to be misclassified.



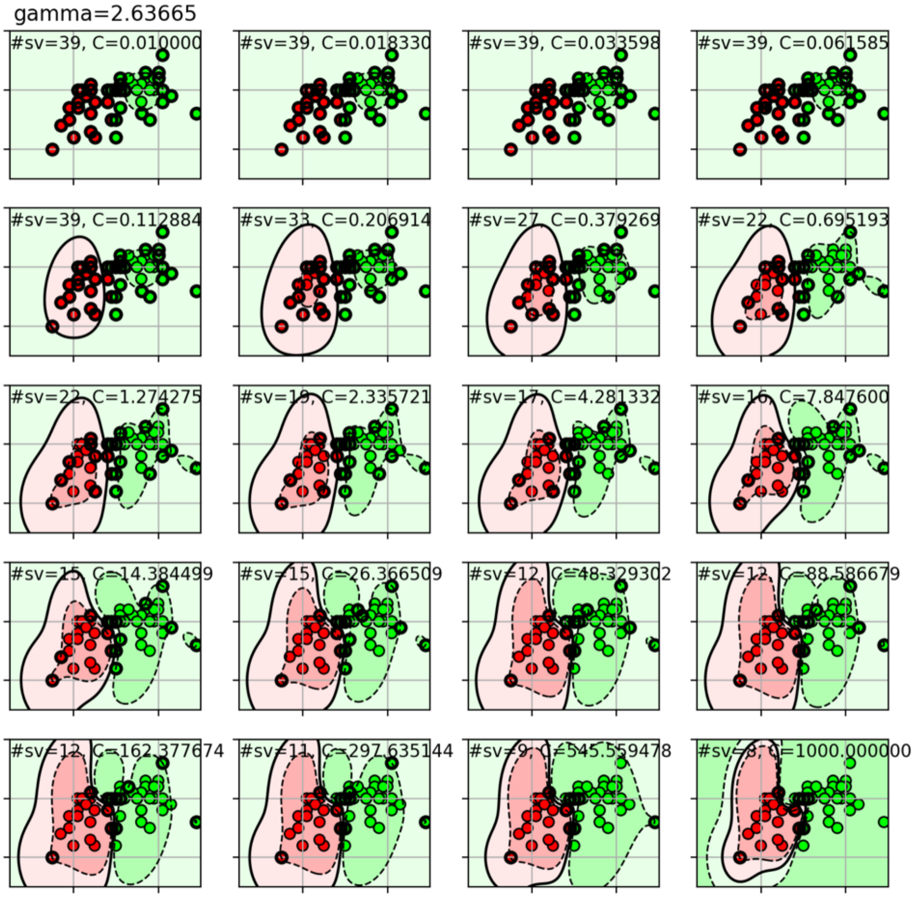
(c) [2 marks] As for the situation in the figure, model with $C \rightarrow 0$ is better for test, since trusting some specific samples may cause the model influenced by noise, while maximizing the margin between sample groups is conducive to the model to not overfit the training data.

Q10. Consider a binary classification problem where the input feature space is d=1,000 dimensions, and we have 2,000 training samples. Rank the following classifiers based on their memory usage (from lowest to highest) when implementing their respective *prediction* functions: linear SVM with C=0.1, linear SVM with C=1000, RBF kernel SVM with C=0.1, RBF kernel SVM with C=1000. Explain the reasons for your rankings.

[4 marks for ranking; 6 marks for reasons]

1) Linear SVM with C=0.1, Linear SVM with C=1000. For any linear SVM, only the weight vector is required, which is 1000 dimensions + 1 bias.
2) Kernel SVM with C=1000. Memory usage is based on the number of support vectors kept, M. The memory usage is 1000*M + 1, with M>=2. Using smaller C will create more *outlier* support vectors than larger C. In general, all the correctly classified support vectors for larger C will be included as outlier support vectors for a smaller C. So the memory used for larger C is generally less. (here we assume the same gamma parameter is used, but different C values)
3) Kernel SVM with C=0.1. There will be more support vectors because more outliers are allowed.

Here is an incorrect argument: when the boundary is very complex, then larger C will have more SV because each point near the boundary (and on the margin) needs to be an SV. Smaller C will have fewer SV because the boundary is less complex. This is not generally true because **all** the points inside the margin (i.e., outliers) will be SVs. Any SV that was correctly classified when using a larger C will necessarily be a SV for the less complex boundary using smaller C, either as a margin point or as an outlier point that is now misclassified. Besides the original SVs from the larger C, using smaller C will also introduce more margin violations and misclassifications then generally more SVs will occur with smaller C, compared to larger C. Here is an example on the iris data:



Q11. For a random forest *F* of *n* trees, suppose the variance of the error of a decision tree *f* is σ^2, and the correlation coefficient between the errors of two trees is ρ. We will have the error variance of *F* as:

$$var(F) = var\left(\frac{1}{n}\sum_{i=1}^{n} f_i\right) = \frac{1}{n}\sigma^2 + \frac{n-1}{n}\rho\sigma^2$$

(a) Explain the relationship between the number of trees and the correlation coefficient and the ability of *F* to reduce overfitting.

(b) For a fixed number of trees *n*, explain some concrete methods to reduce the error variance of *F*?

(a) [5 marks] As n increases, the contribution of the first term decreases. Meanwhile, as the correlation coefficient decreases, the 2nd term also decreases. Thus, the error variance reduces as n increases and rho decreases. Reducing the error variance, means the classifier is correctly fitting the data, i.e., not overfitting.

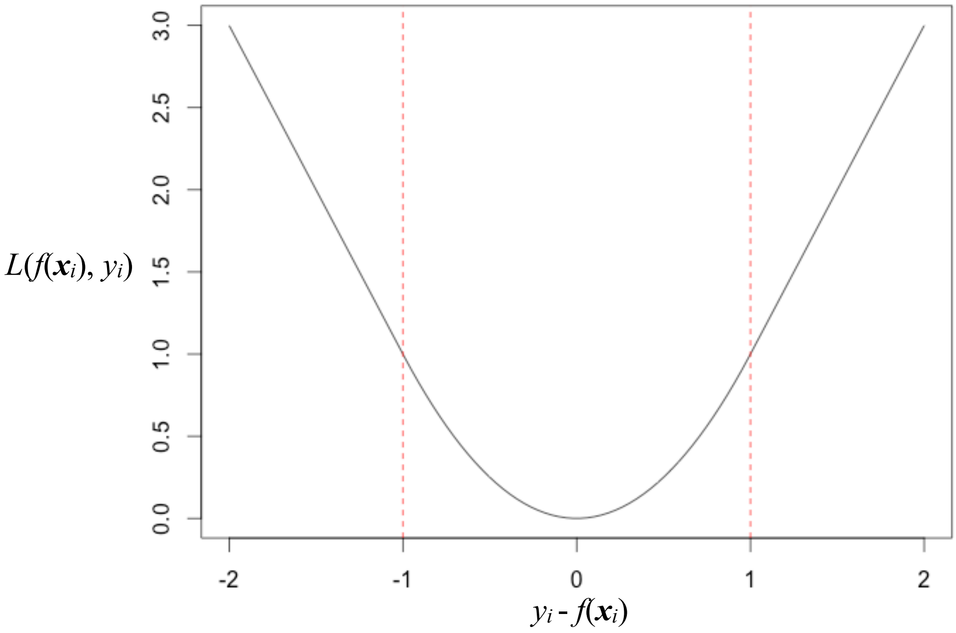
(b) [5 marks] we can train each tree with different subsets of data, and different subsets of features to lower the correlation between the trees.

Q12. Consider linear regression using the Huber loss function:

$$L(f(\boldsymbol{x}_i), y_i) = \begin{cases} \frac{1}{2}(y_i - f(\boldsymbol{x}_i))^2, & \text{if } |y_i - f(\boldsymbol{x}_i)| < 1 \\ |y_i - f(\boldsymbol{x}_i)| - \frac{1}{2}, & \text{otherwise} \end{cases}$$

(a) Draw a plot of the loss function.
(b) Explain one benefit of using Huber loss for linear regression.

(a) [5 marks]



b) [marks] The L1 loss on large values does not prefer to reduce these large errors. Thus the regressor will be robust to outlier points, compared to standard least-squares (L2 loss) regression.

Between errors of -1 and 1, the regressor will behave like ordinary least-squares.

The loss function is differentiable everywhere, whereas L1 loss is not differentiable at 0.

Q13. You have been asked to build a classifier for an early screening test for diagnosing cancer. After the screening test, a more expensive advanced test will be used to confirm the result. For the screening test, you are given a feature vector, and the goal is to predict whether or not the person has cancer (positive or negative). Since it is a screening test, the classifier should classify all people with cancer as positive, while it is okay for some people without cancer to be misclassified as positive. Discuss two methods for getting this desired behavior from the classifier.

[5 marks each]

- 1) use weights on the classes during training. The points in the positive class should have higher weight than the negative class, so that the classifier focuses more on predicting the positive class correctly.
- 2) change the threshold of the classifier. Usually it is T=0, but we can set it to T<0 to make it predict more positive examples.

--- END ---