

CS534 — Midterm — *Solutions*

Name (Please print):

1. (18 pts) Short questions (please provide short explanations to your answers.)

- a. (4 pts) Considering a training set whose labels were randomly corrupted, for the k -nearest neighbor classifier, which of the following choices of k is more robust to the labeling noise: $k=1$ and $k=3$?

$k=3$ is more robust to the labeling noise because it is less likely that all of the three neighbors' labels were corrupted.

- b. (6 pts) In decision tree learning, a multinomial feature (a categorical feature with more than 2 possible values) can be used to create a multiway split, where each possible value of the feature leads to one branch. Explain why such features tend to be favored when comparing against binary features using information gain. Suggest a solution to the above problem.

Consider the extreme case of a n -valued multinomial feature where each training instance has a distinct values for this feature. Such a multiway split will simply put each training point into its own branch. This will lead to the highest information gain. However, using such split the learned decision tree has no ability to generalize to unseen data points, and will perform poorly on a test set. To correct this bias, one can create a binary feature by testing against a particular feature value, or normalize the information gain by the entropy of that feature.

- c. (4 pts) For a neural network, which of the following choices most affects the trade-off between underfitting (not complex enough to represent the target concept) and overfitting:
- i. The initial weights
 - ii. The learning rate
 - iii. The number of hidden units
 - iv. The choice of batch or online learning algorithms

Answer: iii. The number of hidden units has the highest impact because the more hidden units, the more complex hypotheses the neural net can represent.

- d. (4 pts) In class, we showed that under LDA model (i.e., gaussian class conditional distributions, shared covariance matrix), we can show that $P(y|X)$ can be represented in the form of $\frac{1}{1+\exp(-\theta^T \mathbf{x})}$, where θ is some functions of LDA model parameters. Note that this form is identical to what is used by Logistic regression. Does this mean that both methods make the same modeling assumptions? If your answer is no, which one makes stronger modeling assumptions (more restrictive, harder to satisfy)?

Answer: LDA makes stronger modeling assumptions than LR. LDA assumes that $P(X|y)$ is Gaussian with shared covariance matrix. While this leads to the same $P(y|X)$ form as used by LR, other probabilistic models can also lead to the same $P(y|X)$ form. Thus LDA poses stronger modeling assumption.

2. (20 pts) Naive Bayes Classifier. Consider the following training data set.

A	B	C	Y
0	1	1	0
1	1	1	0
0	0	0	0
1	1	0	1
0	1	0	1
1	0	1	1

- a. (6 pts) Write down the Naive Bayes Classifier that would be learned on this data set assuming no laplace smoothing. In particular, write down the values of each learned parameter.

$$P(y = 1) = 0.5$$

$$P(A = 1|y = 1) = 2/3, P(A = 1|y = 0) = 1/3$$

$$P(B = 1|y = 1) = 2/3, P(B = 1|y = 0) = 2/3$$

$$P(C = 1|y = 1) = 1/3, P(C = 1|y = 0) = 2/3$$

- b. (5 pts) What predictions will the Naive Bayes Classifier make for (A=1, B=0, C=0)? Show your work.

$$P(y = 1|A = 1, B = 0, C = 0)$$

$$\begin{aligned} &\propto P(y = 1)P(A = 1|y = 1)P(B = 0|y = 1)P(C = 0|y = 1) \\ &= \frac{1}{2} \times \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} = \frac{2}{27} \end{aligned}$$

$$P(y = 0|A = 1, B = 0, C = 0)$$

$$\begin{aligned} &\propto P(y = 0)P(A = 1|y = 0)P(B = 0|y = 0)P(C = 0|y = 0) \\ &= \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{54} \end{aligned}$$

Predict $y=1$.

- c. (4 pts) What is the potential problem with using a Naive Bayes Classifier that was learned without Laplace smoothing?

When we have limited training data, the maximum likelihood estimation of some of the probabilities may take value zero, causing the posterior probability to be zero for both $y=1$ and $y=0$.

- d. (5 pts) Suppose now we are told that $Y = B \text{ XOR } C$. Do you think the Naive Bayes Classifier is appropriate for this problem? Why? *No. Because this indicates that the naive bayes assumption is violated, i.e., B and C are not conditionally independent given y .*

3. [10 pts] **LOO parameter selection for K-NN.** Consider the following one-dimensional data set, where we use 'x' and 'o' to denote different classes.



- a. [4 pts] What is the training error of 1-Nearest Neighbor?

The training error of 1-NN is zero.

- b. [6 pts] Apply leave-one-out cross validation to choose between $k = 1$ and $k = 3$ for k-Nearest Neighbor.

For $k=1$, we make the following predictions: x, x, x, x, x, o, x, o. The 4-7th points are misclassified. Thus the LLO error rate for $k=1$ is 50%. For $k=3$, we make the following predictions: x, x, x, x, o, o, o, o. The 4th and the 6th points are misclassified. Thus $k=3$ has LLO error rate of 25%. So we choose $k=3$.

4. [20 pts] (Support vector machines)

- a. (5 pts) Please provide the mathematical definition of the functional and geometric margins for linear SVM. Explain why geometric margin is a better objective function for learning maximum margin classifier.

Functional margin: $y(\mathbf{w}\mathbf{x} + b)$

Geometric margin: $\frac{y(\mathbf{w}\mathbf{x}+b)}{\|\mathbf{w}\|}$

Geometric margin is more appropriate objective ftn because functional margin can be scaled arbitrarily by scaling \mathbf{w} and \mathbf{x} without changing the decision boundary.

- b. (5 pts) Consider the objective function that the softmargin SVM tries to minimize $\|\mathbf{w}\|^2 + C \sum_i \xi_i$. Please explain why softmargin SVM can be interpreted as performing Structural risk minimization.

SRM aims to minimize the upper error bound, which consists of the training error and the structural penalty based on the vc-dimension. Similarly, Softmargin SVM tradeoffs between minimizing the sum of the slack variable, which can be viewed as a measure of training error, and maximizing the margin, which reduces the effective vc-dimension of the hypothesis space.

- c. (5 pts) Please give a 1-d example where the data is linearly separable, yet you expect the softmargin SVM to produce a classifier with non-zero training error.

Consider the following examples:

xxxx o oooooo

It is linearly separable. However, if we set c to be small enough for softmargin SVM, we will likely treat the leftmost 'o' point as an outlier and produce a decision boundary that will misclassify that point in order to achieve a larger margin.

- d. (5 pts) Please describe how the kernel trick (i.e., using kernel functions in place of inner product) can address the linearly non-separability issue for SVM without incurring significant additional computational cost.

By using the kernel function, we are performing an implicit mapping from the original input space to a potentially non-linear feature space, making data linearly separable in this mapped feature space, thus solving the linearly non-separability issue for SVM. Further, using the kernel function does not require explicitly computing this mapping, thus incurs no significant computational burden.

5. (10 pts) VC-dimensions

- a. (4 pts) Consider decision trees with real-valued features, what is the VC-dimension of the hypothesis space containing unbounded decision trees. Explain your answer.

Infinite. This is because H can shatter arbitrary number of data points since using unbounded decision trees we can also achieve zero training error on any arbitrary training set.

- b. (6 pts) Prove that $VC(H) \leq \log_2 |H|$

Let m be H 's vc-dimension. That is to say there exists a set of size m that can be shattered by H . Given m data points, there are 2^m possible ways to label them. To shatter these m points, we need at least 2^m hypotheses in H . This suggests that $|H| \geq 2^m$. As a result, we have $m = VC(H) \leq \log_2 |H|$

6. [20 pts] PAC-learning

- a. (10 pts) We have shown in class that if a consistent learning algorithm for a finite hypothesis space is provided with

$$m \geq \frac{1}{\epsilon} (\ln |H| + \ln \frac{1}{\delta})$$

randomly drawn training examples, then we can state a certain guarantee. Please clearly state this guarantee. Be sure to explain the roles of ϵ and δ .

The guarantee states that with at least $1 - \delta$ probability that the consistent learning algorithm will output a hypothesis whose generalization error is no greater than ϵ .

- b. (10 pts) Consider the concept class C of boolean conjunctions over n boolean variables. In other words, the input space contains n boolean features, and the concept class C contains all possible conjunctions over these boolean variables. Prove that C is PAC-learnable. You can use the above inequality if you would like.

See hw4 problem 10 solution.