Theory Assignment 1

COMP 451 - Fundamentals of Machine Learning

Prof. William L. Hamilton

Winter 2021

Preamble The assignment is due February 9th at 11:59pm via MyCourses. Late work will be automatically subject to a 20% penalty, and can be submitted up to 5 days after the deadline. You may scan written answers or submit a typeset assignment, as long as you submit a single pdf file with clear indication of what question each answer refers to. You may consult with other students in the class regarding solution strategies, but you must list all the students that you consulted with on the first page of your submitted assignment. You may also consult published papers, textbooks, and other resources, but you must cite any source that you use in a non-trivial way (except the course notes). You must write the answer in your own words and be able to explain the solution to the professor, if asked.

Question 1 [6 points]

Recall that the k-NN model is defined by the prediction function

$$f_{\text{k-NN}}(\mathbf{x}) = \texttt{MAJ}\left(\{y_i : (\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{trn}} \wedge \exists_{<k}(y_j, \mathbf{x}_j) \in \mathcal{D}_{\text{trn}} : d(\mathbf{x}, \mathbf{x}_i) > d(\mathbf{x}, \mathbf{x}_j)\}\right), \tag{1}$$

where MAJ is the majority vote function. Assume that we are considering a binary 0-1 classification task with two-dimensional features, and that we are evaluating accuracy using the 0-1 loss:

$$L(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

Lastly, assume that ties in the majority vote function are broken randomly with a 50/50 probability, and assume that we are evaluating the expected accuracy in light of this randomness.

Prove or provide a counter-example to the following claim: the prediction error (i.e., the 0-1 loss) on the training set for a k-NN with $k = |\mathcal{D}_{\text{trn}}|$ is always at least as large as for the training error for a model with $k = 1$, and there exists a training set where these two models have identical training error.

Hint: Remember that the nearest neighbor of a training point is always itself!

Question 2 [6 points]

For this question, you should refer to the details and notation for the perceptron algorithm (i.e., Algorithm 1) in Chapter 3 of the notes. Provide a proof for the following lemma, which we we used to prove the perceptron convergence theorem:

Lemma 1. *Assume that there exists some $\gamma > 0$ and some set of optimal parameters $\mathbf{w}^*$ such that $y_i(\mathbf{w}^*)^\top \mathbf{x}_i \geq \gamma$ for all $(\mathbf{x}_i, y_i) \in \mathcal{D}_{trn}$. Then we have that the inner product $(\mathbf{w}^*)^\top \mathbf{w}^{(k)}$ increases at least linearly with each update in the perceptron algorithm. In particular, we have that $(\mathbf{w}^*)^\top \mathbf{w}^{(k)} \geq \gamma k$, where k denotes the number of updates in Algorithm 1.*

Question 3 [6 points]

In class, we were introduced to Bernoulli Naive Bayes and the Gaussian Naive Bayes models. In this question, you will derive that maximum likelihood parameters for a Binomial Naive Bayes model, which could be used for count-based data. Here we assume that $\mathbf{x}[j] \in \{0, 1, ..., N\}$ meaning that the feature vector contains positive integers with a maximum value of N. For example, in text classification, each feature could count how often a particular word occurs in a document, and we assume that the documents all have a maximum length of mN. In this Binomial Naive Bayes model, the feature likelihood is defined following distribution:

$$P(\mathbf{x}[j] \mid y = k) = \binom{N}{\mathbf{x}[j]} \theta_{j,k}^{\mathbf{x}[j]} (1 - \theta_{j,k})^{N - \mathbf{x}[j]} \tag{3}$$

As in the Bernoulli Naive Bayes model, the $\theta_{j,k}$ parameter determines the likelihood of seeing the jth feature. However, in this case, the outcome is a count from 0 to N, rather than a binary value.

Part 1 [2 points]

Assume we are in a binary classification setting. Write an expression for the log-odds ratio of the Binomial Naive Bayes model. Use the notation from Equation 3 above and use $\theta_1 = P(y = 1)$ to denote the estimated class likelihood.

Part 2 [4 points]

Derive the maximum likelihood estimates for the Binomial Naive Bayes parameters, i.e., give maximum likelihood estimates for the $\theta_{j,k}$ parameters.

Question 4 [short answers; 2 points each]

Answer each question with 1-3 sentences for justification, potentially with equations/examples for support.

a) Which of the following statements is false:

1. Generative classification models are always better than discriminative classification models on small datasets.
2. Gaussian Naive Bayes assumes conditional independence between the feature likelihoods.
3. The perceptron algorithm is guaranteed to converge on linearly separable datasets.
4. A 1-NN model can misclassify some training points.
5. Bernoulli Naive Bayes is only well-defined for binary features.

b) Consider the following dataset:

point 1:	$([0, 1], 1)$
point 2:	$([0, 1], -1)$
point 3:	$([0, 1], 1)$
point 4:	$([0, 1], -1)$

Assume we run the perceptron algorithm (i.e., Algorithm 1 from Chapter 3) with initial guess $\mathbf{w}^{(0)} = [0, 0], b^{(0)} = 0$ and that we iterate over the data in the order given above. What would be the parameter values $\mathbf{w}^{(100)}$ and $b^{(100)}$ *after* the 100th update?

c) Consider the following training dataset:

point 1:	$([0,1],1)$
point 2:	$([1,0],1)$
point 3:	$([0,0],0)$
point 4:	$([1,1],0)$

What probability would a Bernoulli Naive Bayes model assign to point 4 belonging to class 1? In other words, what is $P(y=1 \mid [1,1])$ for a Bernoulli Naive Bayes Model trained on this data.