Assignment 2

CMPT 727
Spring 2020

Problem 1

Given that we wish to learn some target function $f : X \longrightarrow Y$, give a brief description on how to apply Bayes rule as the basis for designing learning algorithms (generative classifier).

Problem 2

1. Given $X \perp\!\!\!\perp Y|Z$, can we say $P(X,Y|Z) = P(X|Z)P(Y|Z)$? Explain
2. Given $X \perp\!\!\!\perp Y|Z$, can we say $P(X,Y) = P(X)P(Y)$? Explain.
3. Suppose X is a vector of n boolean attributes and Y is a single discrete-valued variable that can take on J possible values. Let $\theta_{ij} = P(X_i|Y = y_j)$. What is the number of independent θ_{ij} parameters?

Problem 3 (Textbook exercise 3.1): MLE for the Bernoulli/ binomial model

Derive Equation $\hat{\theta}_{MLE} = \frac{N_1}{N}$ by optimizing the log of the likelihood in Equation $p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$.

Problem 4 (Textbook exercise 3.6): MLE for the Poisson distribution.

The Poisson distribution pmf is defined as

$$\text{Poi}(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!},$$

for $x \in \{0, 1, 2, \dots\}$ where $\lambda > 0$ is the rate parameter. Derive the MLE.

Problem 5 (Textbook exercise 3.12): MAP estimation for Bernoulli with non-conjugate priors

Suppose we flip a coin N times and observe N_0 tails and N_1 heads. Consider the following prior that believes the coin is either fair or slightly biased towards tails.

$$p(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.5 \\ 0.5 & \text{if } \theta = 0.4 \\ 0 & \text{otherwise} \end{cases}$$

1. Derive the MAP estimate under this prior.
2. Suppose the true parameter is $\theta = 0.41$. When N is small, will this prior or a Beta$(1, 1)$ prior lead to a MAP estimate that is closer to the true θ? What about when N is large?

Problem 6 (Textbook exercise 3.21). Feature selection using a naive Bayes classifier with binary features.

One way to select features for a naive Bayes classifier is to use mutual information between each feature X_j and the class label Y:

$$I(X_j, Y) = \sum_{x_j} \sum_{y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)}.$$

Let $\pi_c = p(y = c)$, $\theta_{jc} = p(x_j = 1|y = c)$ and $\theta_j = p(x_j = 1) = \sum_c \pi_c \theta_{jc}$. Show that the MI can be computed as follows

$$I(X_j, Y) = \sum_c \left[\theta_{jc}\pi_c \log \frac{\theta_{jc}}{\theta_j} + (1 - \theta_{jc})\pi_c \log \frac{1 - \theta_{jc}}{1 - \theta_j} \right].$$

Problem 7 (Textbook exercise 3.20): Comparing naive Bayes with full conditional model

Consider a generative classifier for C classes with class conditional density $p(\mathbf{x}|y)$ and uniform class prior $p(y)$. Suppose all the D features are binary $x_j \in \{0, 1\}$. If we assume all the features are conditionally independent (the naive Bayes assumption), we can write

$$p(\mathbf{x}|y = c) = \prod_{j=1}^{D} \mathrm{Ber}(x_j|\theta_{jc}0$$

This requires DC parameters.

1. Now consider a different model, which we will call the "full" model, in which all the features are fully dependent (i.e. we make no factorization assumptions). How might we represent $p(\mathbf{x}|y = c)$ in this case? How many parameters are needed to represent $p(\mathbf{x}|y = c)$?
2. Suppose we train each model by finding the MLE on a training set of N cases. If the sample size N is very small, which model (naive Bayes or full) is likely to give lower test set error, and why?
3. If the sample size N is very large, which model is likely to give lower test set error, and why?
4. What is the computational complexity (in big-Oh notation) of fitting the full and naive Bayes models respectively as a function of N and D? (You may assume you can convert a D-bit array to an array index in $O(D)$ time.)
5. What is the computational complexity of applying the full and native Bayes models respectively at test time to a single test case?
6. Suppose the test case has missing data. Let $\mathbf{x}_v$ be the visible features of size v and let $\mathbf{x}_h$ be the hidden (missing) features of size h, where $v + h = D$. What is the computational complexity of computing $p(y|\mathbf{x}_v, \hat{\theta}$ for the full and naive bayes models, as a function of v and h?

Problem 8 (Textbook exercise 5.3). Reject option

In many classification problems one has the option either of assigning x to class j or, if you are too uncertain, of choosing the reject option. If the cost for rejects is less than the cost of falsely classifying the object, it may be the optimal action. Let α_i mean you choose action i, for $i = 1 : C + 1$, where C is the number of classes and $C + 1$ is the reject action. Let $Y = j$ be the true (but unknown) state of nature. Define the loss function as follows:

$$\lambda(\alpha_i|Y = j) = \begin{cases} 0, & \text{if } i = j \text{ and } i, j \in \{1, ..., C\} \\ \lambda_r, & \text{if } i = C + 1 \\ \lambda_s, & \text{Otherwise} \end{cases}$$

In other words, you incur 0 loss if you correctly classify, you incur λ_r loss (cost) if you choose the reject option, and you incur λ_s loss (cost) if you make a substitution error (misclassification).

- Show that the minimum risk is obtained if we decide $Y = j$ if $p(Y = j|x) \geq p(Y = k|x)$ for all k (i.e., j is the most probable class) and if $p(Y = j|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$; otherwise we decide to reject.)

- Describe qualitatively what happens as λ_r/λ_s is increased from 0 to 1 (i.e., the relative cost of rejection increases).