CMSC 35900-2: A Probabilistic Approach to Machine Learning

Problem set 1

Due Thursday, September 28th

Exchangeability

Recall the definition of an exchangeable sequence of random variables:

Definition 1. *The random variables $X_1, X_2, \ldots, X_N$ are* **exchangeable** *iff for every permutation $\pi \in \mathbf{S}_N$, the joint distribution of $(X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(N)}$ is identical to the joint distribution of $X_1, X_2, \ldots, X_N$.*

Definition 2. *The infinite sequence of random variables $X_1, X_2, \ldots$ is* **exchangeable** *iff every finite subset of the variables is exchangeable.*

de Finetti's Theorem states that an *infinite* sequence of random variables is exchangeable iff the random variables are i.i.d. conditioned on some other random variable.

Certainly if $X_1, X_2, \ldots$ are i.i.d. given Z than they are exchangeable, regardless of whether the number of exchangeable random variables is finite or infinite. But the converse is not necessarily true if the sequence is infinite.

Problem 1 Give a concrete joint distribution over two random binary random variables X_1, X_2 such that X_1, X_2 are exchangeable but there exists no Z such that they are i.i.d. given Z.

Model Comparison for the Naïve Bayes

In this section we consider two alternate Naïve Bayes models, H_2 based on only two features, and H_3 based on three features. As in class, we use Y_i to denote the binary label, and $X_i[1], \ldots, X_i[3]$ to denote the binary features, with:

$$Y_i|q \sim \text{Ber}(q) \qquad q \sim \text{Uniform}[0,1] X_i[j]|Y, p \sim \text{Ber}(p^{Y_i}[j]) \qquad p^y[j] \sim \text{Beta}(\alpha, \alpha). \tag{1}$$

With α a fixed parameter. In H_3 the conditional distributions (1) are for $j = 1, 2, 3$, while in H_2 these only describe the first two features, $j = 1, 2$, while for the last feature there is only a single parameter $p[3]$ with:

$$X_i[3]|p \sim \text{Ber}(p[3]) \qquad\qquad p[3] \sim .\text{Beta}(\alpha, \alpha) \tag{2}$$

Notice that both models describe a joint distribution for $(Y_i, X_i[1], X_i[2], X_i[3])$, but in H_2 the third feature is completely independent of the label and the two other features.

Our goal is to choose between the two models by comparing the Bayesian evidence for each one. To do so, it will be useful to first remind ourselves of the evidence for a simple Beta-Bernoulli model:

Problem 2 Let $\theta \sim \text{Beta}(\alpha, \beta)$ and $Z_i|\theta \sim \text{i.i.d.Ber}(\theta)$. Calculate the Bayesian evidence $P(Z_1, \ldots, Z_n)$ for a sequence of n observations $Z_1, \ldots, Z_n$, k of which are positive.

Using the above calculation, we now turn to comparing the two models H_2 and H_3. We will use $D = ((Y_1, X_1[2], X_1[2], X_1[3]), \ldots, (Y_N, X_N[2], X_N[2], X_N[3]))$ to denote an observed training set of N labeled examples. We will also refer to the following sets of random variables: $Y = (Y_1, Y_2, \ldots, Y_N)$ and $X[j] = (X_1[j], X_2[j], \ldots, X_N[j])$.

Problem 3

1. First consider the likelihood of the maximum-likelihood estimates under each of the models. Which one will *always* be higher (or equal)? Why?
2. Using the Bayes Ball rules or the notion of d-separation, explain why, in each of the two models, for each $j \neq j'$ we have $X[j] \perp X[j']|Y, H$ (recall each of these refers to a *set* of random variables, as defined above).
3. Conclude that in each of the two models the Bayesian evidence factorizes as: $P(D|H) = P(Y|H) \prod_j P(X[j]|Y, H)$.
4. Write down an expression for $P(X[3]|Y, H)$ under each of the two models, in terms of counts of the form $\sharp(Y = y, X[j] = x)$.
5. Write down an expression for the evidence ratio $\frac{P(D|H_2)}{P(D|H_3)}$.

Naïve Bayes with Continuous Features and the Gaussian Mixture Model

We would now like to consider a model with the same dependency structure as the Naïve Bayes model we studied in class, but where $X_i[j]$ are real-valued (instead of binary) and Gaussian distributed conditioned on Y_i.

Problem 4 We will first establish that a conjugate prior to the mean of a Gaussian distribution is a Gaussian itself. Let $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and $Z_i|\theta \sim$ i.i.d.$\mathcal{N}(\theta, \sigma)$.

1. Show that the posterior distribution $\theta|Z_1, \ldots, Z_N$ is Gaussian and calculate its mean and variance. I.e. find the values of μ_N, σ_N for which $\theta|Z_1, \ldots, Z_N \sim \mathcal{N}(\mu_N, \sigma_N)$.

2. What are the maximum likelihood, maximum a-posteriori and posterior mean estimates of θ given observations $Z_1, \ldots, Z_N$?

3. Find the posterior distribution of $Z_{N+1}|Z_1, \ldots, Z_N$ (Hint: express Z_{N+1} as $Z_{N+1} = \theta + \mathcal{N}(0, \sigma^2)$). Compare this distribution to just using the posterior mean estimate, i.e. to $P(Z_{N+1}|\theta_{\text{PM}})$.

We now turn the the Naïve Bayes model itself, with real valued features $X_i[j]$. Instead of the parameters $p^y[j]$ we will introduce the parameters $\mu^y[j]$ which are conditional means of $X_i[j]$. The model is then specified as:

$$Y_i|q \sim \text{Ber}(q) \qquad q \sim \text{Uniform}[0,1] \tag{3}$$

$$X_i[j]|Y, p \sim \mathcal{N}(\mu^{Y_i}[j], \sigma^2) \qquad \mu^y[j] \sim \text{Beta}(\alpha, \alpha). \tag{4}$$

Problem 5

1. Show that, given the parameters, the inverse conditional distribution is of the logistic linear form:
$$P(Y_i|X_i, \mu) = \frac{1}{1 + e^{-(w' X_i + w^0)}}$$
Write down the expression of $w[j]$ and w^0 as a function of the parameters p, q.

2. (Optional) Show that the posterior $P(Y_{N+1}|X_{N+1}, D)$ also has a logistic linear form.

3. (Optional) In this Gaussian Naïve Bayes model, $X_i|Y_i, \mu$ is a spherical Gaussian with fixed variance (only the mean depends on Y_i and μ). That is, considering X_i and μ^y as a vector in $\mathbb{R}^k$, we have $X_i|Y_i, \mu \sim \mathcal{N}(\mu^{Y_i}, \sigma^2 I)$. Now consider a more general model in which the variance is not fixed, but rather $\sigma^0, \sigma^1 \in \mathbb{R}$ are parameters and $X_i|Y_i, \mu \sim \mathcal{N}(\mu^{Y_i}, (\sigma^{Y_i})^2 I)$. What form does $Y_i|X_i, \mu, \sigma$ take under this model? What are the possible decision boundaries for prediction using estimated values of parameters? How do the answers change when the Gaussian are not restricted to be spherical: $X_i|Y_i, \mu \sim \mathcal{N}(\mu^{Y_i}, \Sigma^{Y_i})$ with $\Sigma^0, \Sigma^1 \in \mathbb{R}^{k \times k}$ being arbitrary covariances matrices?

Naïve Bayes vs. Linear Discrimination

We saw how for both the Bernoulli Naïve Bayes model and the Gaussian Naïve Bayes model the inverse conditional $Y_i|X_i$, params is a linear logistic and the decision boundary is linear. We will now explore the differences between the two models.

Problem 5 We first consider a Gaussian Naïve Bayes model (as in equations (**??**)) with a single feature $X_i[1]$ and with $\sigma^2 = 1$ and $\sigma_0^2 = 10^6$ (this corresponds to a fairly weak prior on μ), and a linearly separable data set consisting of the four labeled points:

$$(X_1 = -1, Y_1 = 0), (X_2 = -1, Y_1 = 0), (X_3 = 1, Y_3 = 1), (X_4 = 100, Y_4 = 1)$$

1. What is the decision boundary of discriminative logistic regression trained by maximizing the conditional likelihood on this data set? Imposing a proper prior on the weight vector of such a logistic model would change the decision boundary slightly, but with a weak enough prior (high enough variance on the weight vector), this will be a very small change. What is the training error for such a predictor (i.e. how many errors will such a predictor make on the training set itself)?

2. What is the decision boundary of a predictor obtained by using the maximum likelihood parameter settings $\mu_{\text{MAP}}, q_{\text{MAP}}$ of the Gaussian Naïve Bayes classifier? Using the true MAP parameter settings, or the true posterior, would yield very similar predictions since the prior $\mu^y \sim \mathcal{N}(0, 10^6)$ is very weak. What is the training error for such a predictor?

3. Why does a generative approach fail here?

4. (Optional) Construct a data set displaying a similar behavior for the Bernoulli Naïve Bayes model. This time one feature will not be enough, but it is possible to construct an example with four features, such that the training set is linearly separable while the predictor obtained by using the maximum likelihood parameter setting (nor using the MAP setting or the posterior mean) of the Naïve Bayes model does not correctly separate the training data.