EE5907/EE5027 Week 3: Univariate Gaussian + Naive Bayes

Some of the following questions are adapted from Kevin Murphy's (KM) book "Machine Learning: A Probabilistic Perspective".

Q1: Mixed Observations Naive Bayes

Consider a 3-class naive Bayes classifier with one binary feature and one Gaussian feature. More specifically, class label y follows a categorical distribution parametrized by π, i.e., $p(y = c) = \pi_c$. The first feature x_1 is binary and follows a Bernoulli distribution: $p(x_1|y = c) = \text{Bernoulli}(x_1|\theta_c)$. The second feature x_2 is univariate Gaussian: $p(x_2|y = c) = \mathcal{N}(x_2|\mu_c, \sigma_c^2)$. Let $\pi = [0.5\ \ 0.25\ \ 0.25]$, $\theta = [0.5\ \ 0.5\ 0.5]$, $\mu = [-1\ \ 0\ \ 1]$ and $\sigma^2 = [1\ \ 1\ \ 1]$.

(i) Compute $p(y|x_1 = 0)$. Note that result is a vector of length 3 that sums to 1.

(ii) Compute $p(y|x_2 = 0)$. Note that result is a vector of length 3 that sums to 1.

(iii) Compute $p(y|x_1 = 0, x_2 = 0)$. Note that result is a vector of length 3 that sums to 1.

Exercise 3.20 Class conditional densities for binary data

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 native Bayes assumption), we can write

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

This requires DC paramters.

a. 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 $p(\mathbf{x}|y = c)$?

b. Assume the number of features D is fixed. Let there be N training 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?

c. If the sample size N is very large, which model (naive Bayes or full) is likely to give lower test set error, and why?

d. What is the computational complexity of fitting the full and naive Bayes models as a function of N and D? Use big-Oh notation. (Fitting the model here means computing the MLE or MAP parameter estimates. You may assume you can convert a D-bit vector to an array index in $O(D)$ time.)

e. What is the computational complexity of applying the full and naive Bayes models at test time to a single test case?

f. Suppose the test case has missing data. Let $\mathbf{x}_v$ be the visible features of size v, and $\mathbf{x}_h$ 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?

Q3: Posterior Predictive Distribution for Exponential Distribution

(a) Consider an exponential distribution $p(x) = \lambda e^{-\lambda x}$. Suppose we observe N independent samples from the exponential distribution: $D = \{x_1, \cdots, x_N\}$.

- (i) What is the maximum likelihood (ML) estimate of λ? Show your steps to get full credit.
- (ii) Suppose we use ML estimate of λ to predict new data x_{N+1}. What problems might arise? Describe a solution to avoid this problem.

(b) Consider the same distribution and data from part (a). Assume the conjugate prior distribution $p(\lambda) = \text{Gamma}(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta}$, where $\Gamma(\cdot)$ is the Gamma function (not to be confused with the Gamma distribution). You may or may not find the following identities useful: $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$, and $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$.

- (i) The posterior distribution $p(\lambda|D)$ is also a Gamma distribution with parameters α', β'. What are α' and β'? Show your steps.
- (ii) What is the posterior predictive distribution $p(x_{N+1}|D)$? Show your steps.