CS475 Machine Learning, Fall 2012: Homework 3

Dan Crankshaw - dcranks1

Question 1:

(a) The exponential distribution family takes the form

$$f(X = x; \theta) = h(x)e^{\eta(\theta)T(x) - A(\theta)} \tag{1}$$

(1) The binomial distribution is

$$f(X = x; p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{2}$$

$$= \binom{n}{x} e^{log(p^x(1-p)^{n-x})} \tag{3}$$

$$= \binom{n}{x} e^{xlog(\frac{p}{1-p}) + nlog(1-p)} \tag{4}$$

$$= h(x)e^{\eta(\theta)T(x) - A(\theta)} \tag{5}$$

where $h(x) = \binom{n}{x}$, $\eta(\theta) = log\frac{\theta}{1-\theta}$, $T(x) = x$, and $A(\theta) = -nlog(1-\theta)$.

(2) The Poisson distribution is

$$f(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{6}$$

$$= \frac{1}{x!} e^{xlog\lambda - \lambda} \tag{7}$$

$$= h(x)e^{\eta(\theta)T(x) - A(\theta)} \tag{8}$$

where $h(x) = \frac{1}{x!}$, $\eta(\theta) = log\theta$, $T(x) = x$, and $A(\theta) = \theta$.

(3) The Gaussian distribution is

$$f(X = x; \lambda) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{9}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} e^{\frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}} \tag{10}$$

$$= h(x)e^{\eta(\theta)T(x) - A(\theta)} \tag{11}$$

where $h(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$, $\eta(\theta) = \frac{\mu}{\sigma^2}$, $T(x) = x$, and $A(\theta) = \frac{\mu^2}{2\sigma^2}$.

(b) To be honest, I'm a little unclear at what this question is getting at. Here are some thoughts on the problem though. In our Poisson regression, we need to model $P(Y|X,\theta) = Poisson(x \cdot \theta)$. We know that the mean of the a Poisson Distribution is $\lambda = x \cdot \theta$, so we can say the likelihood function $l(\lambda) = l(x \cdot \theta)$. However, I'm not sure what information $\eta(\theta)$ being the log odds ratio for logistic regression gives us, and what inferences about general exponential family distributions we can make from that.

(c) We still call them linear models because the model itself is linear. It creates a separating hyperplane and is a linear function of x. The non-linearity comes from passing the output of the linear model through a potentially non-linear function (like the logistic function), generally to map the output to a specific range.

(d) If the data are linearly separable, then the maximum likelihood hyperplane occurs when the probability of predicting the correct class for the data goes to 1. That probability is

$$P(Y = y_c|x) = \frac{1}{1 + e^{-w^T \cdot x}} \tag{12}$$

This probability goes to 1 when $w^T \cdot x$ goes to infinity. Because x is fixed, this occurs when w goes to infinity. We can prevent this from happening by adding a regularization term, which is an additional term in the MLE function we are maximizing. This term is a penalty term based on the size of the l_1 or square norm of w, so that the bigger the sizes of the coefficients in w, the larger the penalty incurred. This acts as a counterbalance to the logistic function portion of the equation, which in the case of linearly separable data will tend to go to infinity. This means that the maximum value of the logistic function with regularization will be finite.

Question 2:

(a) There are $2*2*d+2 = 4d+2$ parameters used to describe this mixture. There are 2 paramaters, μ and σ, to describe a univariate Gaussian distribution. We therefore need $2 * d$ parameters to describe d of these distributions, one for each dimension. But because the data comes from a mixture of 2 multivariate Gaussian distributions, each dimension has 2 distributions associated with it, one from each model. That is where the last factor of 2 comes from. The extra 2 parameters are the mixing coefficients for each of the two mixtures.

(b) There are d paramaters used in logistic regression, one for each dimension (these are the indices of the weight vector $\mathbf{w}$). There are $2 * d$ parameters used in Naive Bayes ($P(X|Y = 0)$ and $P(X|Y = 1)$). Naive Bayes makes the assumption that the features in x are conditionally independent.

(c) Generative models maximize likelihood, whereas discriminative models maximize conditional likelihood.

Question 3: Given a trained naive Bayes classifier, if we encounter a missing value for a feature, we can ignore that feature when making our probability predictions. This is possible because of the conditional independence assumption made for naive Bayes models. This means that

$$P(Y, X_1, X_2, \ldots, X_{d-1}) = P(Y)\Pi_{j=1}^{d-1} P(X_i|Y)_{i=1}^{K} P(X_{d,i}|Y) \tag{13}$$

$$= P(Y)\Pi_{j=1}^{d-1} P(X_i|Y)1 \tag{14}$$

Not knowing the value of one feature has no effect on the conditional probabilities of the other features, and because the sum of the conditional probabilities of all possible values for the missing feature is 1, we can ignore that feature.

Question 4: λ helps to increase the bias of the model and decrease the variance. Add-λ smoothing gives some small probability to every event and prevents any zero conditional probabilities from appearing in our model, no matter what the training data is. By giving some small probability to every event (instead of 0 probability), we lessen the amount that the model will vary based on the training data, preventing us from over-fitting any particular training model. This is biasing the model in favor of smoother or more uniform probability distributions.