CS281 Section #3 Notes, v 1.0

Due: never

1 Exponential Family

1.1 Review:

$$p(x|\theta) = \frac{1}{Z(\eta(\theta))}h(x)e^{\eta(\theta)^T\Phi(x)} = h(x)e^{\eta(\theta)^T\Phi(x)-A(\eta(\theta))}$$

η is called **natural or canonical parameters**; $Z(\eta) = \int h(x)e^{\eta(\theta)^T\Phi(x)}dx$ is called **partition function** $A(\eta)$ is called **log-partition function** or **cumulant function**; and $\Phi(x)$ is sufficient statistics of x.
Properties of exponential family are:

- Derivatives of $A(\eta)$ gives cumulants of the sufficient statistics; and $A(\eta)$ is a convex function.
- In canonical exponential families ($\eta(\theta) = \theta$), the log-likelihood function has at most one local maximum (thus it's a global maximum) and MLE must satisfy the empirical average of the sufficient statistics equal the expectation of sufficient statistics.

Conjugate priors exist for exponential family. Likelihood is:

$$p(D|\theta) \propto e^{\eta(\theta)^T\sum_{i=1}^N\Phi(x_i)-NA(\eta(\theta))} = e^{\eta(\theta)^TN\bar{s}-NA(\eta(\theta))}$$

the conjugate prior takes the form:

$$p(\theta|N_0, s_0) \propto e^{\eta(\theta)^TN_0s_0-N_0A(\eta(\theta))}$$

Then, the posterior becomes:

$$p(\theta|D, N_0, s_0) = \frac{1}{Z(N_0+N, N_0s_0+N\bar{s})}e^{\eta(\theta)^T(N_0s_0+N\bar{s})-(N_0+N)A(\eta(\theta))}$$

And the posterior predictive density is:

$$\begin{aligned}
p(D'|D, N_0, s_0) &= \int p(D'|\theta)p(\theta|D, N_0, s_0)d\theta \\
&= \prod_{i=1}^{N'}h(x'_i)\frac{1}{Z(N_0+N, N_0s_0+N\bar{s})}\int e^{\eta(\theta)^TN'\bar{s}'-N'A(\eta(\theta))+\eta(\theta)^T(N_0s_0+N\bar{s})-(N_0+N)A(\eta(\theta))}d\theta \\
&= \prod_{i=1}^{N'}h(x'_i)\frac{Z(N_0+N+N', N_0s_0+N\bar{s}+N'\bar{s}')}{Z(N_0+N, N_0s_0+N\bar{s})}
\end{aligned}$$

1.2 Exercises

- Show that Multinomial distribution and Gamma distritbuion are members of the exponential family. What are the sufficient statistics and natural parameters in each case?
- Check that the first and second derivatives of $A(\eta)$ is the mean and variance of sufficient statistics in each cases above.
- Check that at MLE, the moment equations are satisfied.
- Derive the conjugate prior for Poisson distribution in exponential family form, its posterior and posterior predictive distribution.

1.3 Solutions

- For multinomial distribution with K categories and N trials,

$$\begin{aligned} p(x_1, ...x_K|p_1, ..., p_K) &= \frac{N!}{\prod_{k=1}^K x_k!} \prod_{k=1}^K p_k^{x_k} \\ &= \frac{N!}{\prod_{k=1}^K x_k!} e^{\sum_{k=1}^K x_k \log(p_k)} \\ &= \frac{N!}{\prod_{k=1}^K x_k!} e^{\sum_{k=1}^{K-1} x_k \log \frac{p_k}{1-\sum_{j=1}^{K-1} p_j} + N \log(1-\sum_{j=1}^{K-1} p_j)} \end{aligned}$$

  so $\Phi(x) = \{x_1, ..., x_{K-1}\},\ h(x) = \frac{N!}{\prod_{k=1}^K x_k!},\ \eta(p) = \{\log \frac{p_k}{1-\sum_{j=1}^{K-1} p_j}, k = 1, 2, ..., K-1\},\ A(p) = -N \log(1 - \sum_{j=1}^{K-1} p_j)$

  For Gamma distribution,

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} = e^{(\alpha-1)\log(x) - \beta x - (\log \Gamma(\alpha) - \alpha \log(\beta))}$$

  so $\Phi(x) = (\log(x), x),\ h(x) = 1,\ \eta(\alpha, \beta) = (\alpha - 1, -\beta),\ A(\alpha, \beta) = \log \Gamma(\alpha) - \alpha \log(\beta)$

- For multinomial distribution,

$$\begin{aligned} p_k &= \frac{e^{\eta_k}}{1 + \sum_{k=1}^{K-1} e^{\eta_k}} \\ \frac{\partial p_j}{\partial \eta_k} &= p_j \delta_{jk} - p_k p_j \\ \frac{\partial A(\eta)}{\partial \eta_k} &= \sum_{j=1}^{K-1} \frac{\partial A}{\partial p_j} \frac{\partial p_j}{\partial \eta_k} = \frac{N}{1 - \sum_{j=1}^{K-1} p_j} \sum_{j=1}^{K-1} p_j \delta_{jk} - p_k p_j = \frac{N}{1 - \sum_{j=1}^{K-1} p_j} p_k (1 - \sum_{j=1}^{K-1} p_j) = N p_k = E(x_k) \\ \frac{\partial^2 A(\eta)}{\partial \eta_k \eta_j} &= N \frac{\partial p_j}{\partial \eta_k} = N(p_j \delta_{jk} - p_k p_j) = Cov(x_j, x_k) \end{aligned}$$

  For Gamma distribution,

$$\begin{aligned} A(\eta) &= \log \Gamma(\eta_1 + 1) - (\eta_1 + 1) \log(-\eta_2) \\ \frac{\partial A(\eta)}{\partial \eta_1} &= \Psi(\eta_1 + 1) - \log(-\eta_2) = \Psi(\alpha) - \log(\beta) = E(\log(x)) \\ \frac{\partial A(\eta)}{\partial \eta_2} &= -\frac{\eta_1 + 1}{\eta_2} = \frac{\alpha}{\beta} = E(x) \\ \frac{\partial^2 A(\eta)}{\partial \eta_1 \partial \eta_2} &= -\frac{1}{\eta_2} = \frac{1}{\beta} \\ \frac{\partial^2 A(\eta)}{\partial \eta_1^2} &= \Psi'(\alpha) \\ \frac{\partial^2 A(\eta)}{\partial \eta_2^2} &= \frac{\eta_1 + 1}{\eta_2^2} = \frac{\alpha}{\beta^2} = Var(x) \end{aligned}$$

- For multinomial distribution, at MLE $\hat{p}_k = \frac{x_k}{N}$. For Gamma distribution, $\Psi(\hat{\alpha}) - \log(\hat{\beta}) = \overline{\log(x)}$ and $\frac{\hat{\alpha}}{\hat{\beta}} = \bar{x}$.

- For Poisson distribution, $p(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \propto e^{x \log \lambda - \lambda}$, so the conjugate prior would be $p(\lambda|\alpha, \beta) \propto e^{\alpha \log \lambda - \beta\lambda} = \lambda^\alpha e^{-\beta\lambda}$ which is Gamma distribution with parameters $(\alpha + 1, \beta)$. The posterior is also a Gamma distribution with parameters $(\sum_i x_i + \alpha + 1, N + \beta)$. The predictive distribution is

$$p(x'|D, \alpha, \beta) = \frac{1}{x'!} \frac{(N+\beta)^{\alpha + \sum_i x_i + 1} \Gamma(\alpha + \sum_i x_i + x' + 1)}{(N+\beta+1)^{\alpha + \sum_i x_i + x' + 1} \Gamma(\alpha + \sum_i x_i + 1)} \tag{1}$$

$$= \frac{(\alpha + \sum_i x_i + x')!}{x'!(\alpha + \sum_i x_i)!} (\frac{N+\beta}{N+\beta+1})^{\alpha + \sum_i x_i + 1} (\frac{1}{N+\beta+1})^{x'} \tag{2}$$

Here, $Z(\alpha, \beta) = \frac{\Gamma(\alpha)}{\beta^\alpha}$. The above distribution is a negative binomial distribution.

2 Generalized Linear model

2.1 Review:

Generalized Linear model links the mean μ to a linear predictor $X\beta$ such that $g(\mu) = X\beta$. If $g(\mu) = \eta = X\beta$ (natural parameter), g is called **canonical link**. MLE of β is simplified with canonical link. The likelihood equations equate the sufficient statistics $(\sum_{i=1}^n y_i x_{ij}, j = 1, ..., p)$ for the model parameters to their expected values. For GLM with canonical link, the log likelihood is

$$L(\beta) = \sum_{i=1}^n y_i (\sum_j x_{ij}\beta_j) - A(\sum_j x_{ij}\beta_j)$$

then set the derivatives to zero, we will get:

$$\sum_{i=1}^n y_i x_{ij} - x_{ij}\mu_i = \sum_{i=1}^n (y_i - \mu_i) x_{ij} = 0$$

In the canonical link case the log-likelihood is necessarily a concave function, because the log likelihood for an exponential family distribution is concave in the natural parameter. In using iterative methods to find the ML estimates, we do not need to worry about the possibility of multiple maxima for the log-likelihood. For example, for Poisson model, the canonical link is log function, i.e. $\log(Ey_i) = \log(\mu_i) = X\beta$.

2.2 Exercises:

Gamma GLM could deals with data with positive value and having greater variability when mean is greater. Using a gamma GLM with log link function gives similar results to applying a normal linear model to the log-transformed data. Compare Gamma regression with Log-normal regression.
a. Write down Gamma density in terms of mean μ and shape k. Gamma GLMs usually assume k to be constant but unknown, like σ^2 in ordinary linear models, and use either identify or log link to μ.
b. Use the delta method to show that when y has standard deviation proportional to mean (as does the gamma GLM), then log(y) has approximately constant variance (as does the normal linear model), for small standard deviation.
c. The gamma GLM with log link refers to $\log[E(y_i)] = X\beta$, whereas the ordinary linear model for the transformed response refers to $E[\log(y_i)]$. Show that if $\log(y_i)$ has a $N(\mu_i, \sigma^2)$ distribution, then $\log[E(y_i)] = E[\log(y_i)] + \sigma^2/2$.

2.3 Solutions:

(a) $f(y;k,\mu) = \frac{(k/\mu)^k}{\Gamma(k)} e^{-ky/u} y^{k-1}$, $E(y) = \mu$, $Var(y) = \mu^2/k$. The gamma distribution is in the exponential dispersion family with natural parameter $1/\mu$.
(b) Suppose $E(y) = \mu$ and $SD(y) = \sigma = c\mu$. Then we can find the standard deviation of log(y) via the Delta method. Note that $\log(y) \approx \log(\mu) + (y - \mu)\frac{1}{\mu}$. Then $Var(\log(y)) \approx Var(y)/\mu^2 = c^2$. Then, $\log(y)$ has approximately constant variance.
(c) If $\log(y_i) \sim N(\mu_i; \sigma^2)$, then $E(y_i) = e^{\mu_i + \sigma^2/2}$. Thus, $\log(E(y_i)) = \mu_i + \sigma^2/2 = E(\log(y)) + \sigma^2/2$.

3 Review of the naive Bayes classifier

In this section, I try to cover the basics of linear classification with naive Bayes classifier (NBC). I first present the problem of classification as a mathematical one, and then shows how an NBC approaches it.

3.1 The problem

In a typical classification setting, we have N samples to be sorted into C categories. In the spam-email-detection example, N is the number of emails to be sorted, and C is two because there are only two categories, $\{spam, notspam\}$. In addition, for each sample it has J features. There are different ways to generate features for each problem. For example, we can count the occurrence of 500 different words in each email and get $J = 500$, or we could use the average number of words in a sentence and total number of words in the email as features and get $J = 2$. This way, each sample is a J-dimensional vector. Feature selection is an important step in real-life classification problems.

3.2 What's classification?

To classify the samples, we essentially need to figure out given the features of each sample, which category it most likely belongs to. Formally, the label of the n-th sample y_n can be written as

$$y_n = argmax_{y_n} p(y_n|\mathbf{x_n}) = argmax_{y_n} \frac{p(\mathbf{x_n}|y_n)p(y_n)}{Z}, \tag{3}$$

where Z is a normalizer. Since the second expression is given by Bayes' theorem, this is a "Bayes" classifier.

3.2.1 So naive!

The NBC is called "naive" because of the way it estimates $p(\mathbf{x_n}|y_n)$. From basic properties, we know that

$$p(\mathbf{x_n}|y_n) = p(x_{n1}, x_{n2}, ..., x_{nJ}|y_n) \tag{4}$$

This equals to $\prod_j p(x_{nj}|y_n)$ if and only if the J features of $\mathbf{x_n}$ are all independent, which is by no means guaranteed and therefore "naively" assumed.

3.2.2 Supplying the distributions needed

For an NBC to work, it must be supplied with these distributions: $p(x_{nj}|y_n = c)$ for all j, c, and $p(y_n = c)$ for all c. While the latter is usually simply given as a categorical distribution (let's denote the probability of the category c as π_c), the former can take on various forms. In particular, if $p(x_{nj}|y_n = c)$ is assumed to be from a multivariate normal (MVN) distribution, then we can write the MVN distribution as

$$p(\mathbf{x}|y = c) = \mathcal{N}\left(\mu_c, \Sigma^{(c)}_{diag.}\right). \tag{5}$$

The covariance matrix is diagonal (it is only non-zero on the diagonal) because we assumed that different elements have no correlation. The parameters that describe these distributions generally need to be learned from the data.

3.3 Learning the distributions

Denote all the parameters in an NBC as θ. If we have N samples to learn the distributions from, we denote $\mathbf{X}$ as the N-by-J data matrix and $\mathbf{Y}$ as the N sized vector with labels for each sample, then we can write the model likelihood as

$$p\left(\mathbf{X}, \mathbf{Y}|\theta\right) = \prod_{c} \pi_c^{N(\mathbf{Y}=c)} \prod_{j} \prod_{c} \prod_{n, y_n = c} p(x_{nj}|y_n). \tag{6}$$

This seems complicated, so let's explain it a little bit. For each of the C categories, we count how many samples are classified into this category (that's what $N(\mathbf{Y} = c)$ returns). Since for each of these samples, the category has a $p(y_n = c) = \pi_c$, we take the power of this probability. This is the first half of the equation. For the second half, we go to each category, each sample that's classified in this category, and each feature of this sample, and see the conditional. Since all the features and samples are (assumed to be) independent, we take the product of them to get the joint distribution.

Take the logarithm of both sides of eq.6 to obtain the log likelihood form:

$$\log \mathcal{L}(\theta) = \sum_{c} N(\mathbf{Y} = c) \log(\pi_c) + \sum_{j} \sum_{c} \sum_{n, y_n = c} \log p(x_{nj}|y_n). \tag{7}$$

We can learn the parameters by performing maximum likelihood estimation (MLE).

3.4 Finally, can we classify now?

Say we trained our NBC on some data $\mathbf{X}$ to obtain MLE-fitted parameters θ. Now let's try to predict a new sample $x_{n'}$. The probability that it belongs to the category c is given by

$$p(y_{n'} = c|\mathbf{X}, \theta) \propto \pi_c \prod_{j} p(x_{n'j}|y_{n'}). \tag{8}$$

Again, take the logarithm to obtain

$$\log\left[p(y_{n'} = c|\mathbf{X}, \theta)\right] = \log \pi_c + \sum_{j} \log p(x_{n'j}|y_{n'}) + Const. \tag{9}$$

3.5 A special consideration for multinoulli models

To continue with the discussion in lecture, let's consider a specific classification problem. Here, each feature is categorical, and there are only two categories. That is, instead of scalar values, $x_{nj} \in \{0, 1\}$. If this is the case, then we can simply characterize the model with a mean rate μ_{jc}, which corresponds to the probability of each event of the feature j taking place in the category c. An example would be, if the feature in question of an email is whether the tone is positive(0) or negative(1), then μ_{j0} would be probability that a spam email is positive etc. We can thus calculate eq.8

$$p(y_{n'} = c|x_{n'j}) \propto \pi_c \prod_{j} \mu_{jc}^{N(x_j=1)} (1 - \mu_{jc})^{N(x_j=0)}. \tag{10}$$

This is the "informal parameterization" in class. Likewise, we can rewrite eq.9 as

$$\log\left[p(y_{n'} = c|\mathbf{X}, \theta)\right] = \log \pi_c + \sum_{j} \log(1 - \mu_{jc}) + \sum_{j} x_j \log \frac{\mu_{jc}}{1 - \mu_{jc}} + Const. \tag{11}$$

For the purpose of being lazy, we define $\theta_{jc} = \log \frac{\mu_{jc}}{1 - \mu_{jc}}$. To be even lazier, combine all the θ_{jc} and write θ_c. Thus,

$$\sum_{j} x_j \log \frac{\mu_{jc}}{1 - \mu_{jc}} = \theta_{\mathbf{c}}^{\mathbf{T}} \mathbf{X}. \tag{12}$$

We also denote $b_c = \log \pi_c + \sum_j \log(1 - \mu_{jc})$, which does not depend on the data at all. This is how to obtain the final form:

$$p(y_{n'} = c | x_{n'j}) \propto \exp\left(b_c + \theta_{\mathbf{c}}^{\mathbf{T}} \mathbf{X}\right). \tag{13}$$