Franklin Hu, Sunil Pedapudi
CS 194-10 Machine Learning
Fall 2011
Assignment 5

1. Conjugate priors

   (a) Let

$$\text{Likelihood: } \mathbb{P}(x_1,\ldots,x_N) = \prod_i^N \lambda \exp(-\lambda x_i)$$
$$\text{Prior: gamma}(\lambda|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

Then,

Posterior:

$$\begin{aligned}
\mathbb{P}(\lambda|x_1,\ldots,x_N) &= \prod_i^N \lambda \exp(-\lambda x_i)\cdot\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \\
&= \lambda \exp\left(\sum_i^N -\lambda x_i\right)\lambda^{\alpha-1}e^{-\beta\lambda}\frac{\beta^\alpha}{\Gamma(\alpha)} \\
&= \lambda^{\alpha+N-1}\exp\left(\sum_i^N -\lambda x_i - \beta\lambda\right)\frac{\beta^\alpha}{\Gamma(\alpha)} \\
&= \lambda^{\alpha+N-1}\exp\left(-\lambda\sum_i^N x_i + \beta\right)\frac{\beta^\alpha}{\Gamma(\alpha)} \sim \text{gamma}(\alpha+N,\beta+\sum_i^N x_i)
\end{aligned}$$

Since the posterior also has a gamma distribution, we find the updates parameters are of the form $\alpha+N$,$\beta+\sum_i^N x_i$. To find the prediction distribution,

$$\begin{aligned}
\mathbb{P}(x_N+1|x_1,\ldots,x_N) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int \lambda\exp(-\lambda x_{N+1})\cdot\lambda^{\alpha+N-1}\exp\left(-\lambda(\beta+\sum_i^N x_i)\right)d\lambda \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)}\int \lambda\cdot\lambda^{\alpha+N-1}\exp\left(-\lambda(\beta+\sum_i^{N+1} x_i)\right)d\lambda \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)}\int \lambda\cdot P(\lambda|\alpha+N,\beta+\sum_i^{N+1} x_i)d\lambda
\end{aligned}$$

We note that this describes the expectation for λ given a gamma function $\sim \text{gamma}(\lambda|\alpha+N,\beta+\sum_i^{N+1} x_i)$. Therefore,

$$\mathbb{P}(x_N+1|x_1,\ldots,x_N) = \frac{\alpha+N}{\beta+\sum_i^{N+1} x_i}$$

   (b) Given the geometric distribution

$$P(X_i = k|\theta) = (1-\theta)^{k-1}\cdot\theta$$

and the beta distribution

$$\beta(\theta|a,b) = \alpha\theta^{a-1}(1-\theta)^{b-1}$$

we prove that the beta distribution is the conjugate prior for a likelihood with a geometric distribution.

$$\begin{aligned}
P(\theta|X) &= P(\theta) \cdot P(X|\theta) \\
&= \alpha \cdot \theta^{a-1} \cdot (1-\theta)^{b-1} \cdot (1-\theta)^{k-1} \cdot \theta \\
&= \alpha \cdot \theta^{a} \cdot (1-\theta)^{b+k-2} \\
&= \beta(\theta|a+1, b+k-1)
\end{aligned}$$

The posterior has the form of a beta distribution so therefore the beta distribution is the conjugate prior for the geometric distribution.
The update procedure for a beta posterior simply involves updating the a and b parameters

$$\begin{aligned}
a_{N+1} &\leftarrow a_N + 1 \\
b_{N+1} &\leftarrow b_N + k - 1
\end{aligned}$$

2. Bayesian Naive Bayes
3. Logistic regression for credit scoring
   (a) The data structure we chose for logistic regression is simply a class that keeps a set of weights for each of the features, has an update method for updating the weights, and draws predictions using the logit function

$$\text{Probability} = \frac{1}{1+e^{-w^T x}}$$

   (b) The likelihood is

$$\begin{aligned}
L(w) &= \frac{1}{1+e^{-yw^T x}} \\
\text{log likelihood} &= \log \frac{1}{1+e^{-yw^T x}} \\
&= -\log(1+e^{-yw^T x}) \\
\text{negative log likelihood} &= \log(1+e^{-yw^T x})
\end{aligned}$$

Now we compute the gradient of the negative log likelihood

$$\begin{aligned}
\nabla \log(1+e^{-yw^T x}) &= \nabla \log\left(\frac{e^{yw^T x}+1}{e^{yw^T x}}\right) \\
&= \nabla\left(\log(e^{yw^T x}+1) - \log(e^{yw^T x})\right) \\
&= \left(\frac{1}{e^{yw^T x}+1} \cdot e^{yw^T x} \cdot -yx_i\right) - \left(\frac{1}{e^{yw^T x}} \cdot e^{yw^T x} \cdot -yx_i\right) \\
&= yx_i - yx_i \cdot \frac{e^{yw^T x}}{e^{yw^T x}+1} \\
&= yx_i - yx_i \cdot \left(\frac{e^{yw^T x}+1}{e^{yw^T x}}\right)^{-1} \\
&= yx_i - yx_i \cdot (1+e^{-yw^T x})^{-1} \\
&= yx_i\left(1 - \frac{1}{1+e^{-yw^T x}}\right)
\end{aligned}$$

Therefore our update rule is simply

$$\begin{aligned} w_{i+1} &= w_i + \alpha \cdot \nabla L \\ &= w_i + \alpha \cdot y x_i \cdot \left(1 - \frac{1}{1 + e^{-y w^T x}}\right) \end{aligned}$$