CS 194-10, Fall 2011
Assignment 5

1. *Conjugate priors* (30)
   The readings for this week include discussion of *conjugate priors*. Given a likelihood $P(\mathbf{X} \mid \boldsymbol{\theta})$ for a class models with parameters $\boldsymbol{\theta}$, a conjugate prior is a distribution $P(\boldsymbol{\theta} \mid \boldsymbol{\gamma})$ with hyperparameters $\boldsymbol{\gamma}$, such that the posterior distribution

   $$P(\boldsymbol{\theta} \mid \mathbf{X}, \boldsymbol{\gamma}) = \alpha P(\mathbf{X} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta} \mid \boldsymbol{\gamma})$$

   is drawn from the same family, i.e.,

   $$P(\boldsymbol{\theta} \mid \mathbf{X}, \boldsymbol{\gamma}) = P(\boldsymbol{\theta} \mid \boldsymbol{\gamma}') \tag{1}$$

   where $\boldsymbol{\gamma}'$ are the updated hyperparameters. In class we saw that the conjugate prior for the Bernoulli likelihood $P(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}$ is the beta distribution

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

   For data $\mathbf{X}$ containing N_1 1s and N_0 0s, the likelihood is proportional to $\theta^{N_1}(1-\theta)^{N_0}$ and the posterior for θ is $\text{beta}(\theta \mid a+N_1, b+N_0)$.

   (a) Suppose that the likelihood is given by the exponential distribution with rate parameter[1] λ:

   $$P(x_i \mid \lambda) = \lambda e^{-\lambda x_i}\ .$$

   Show that the *gamma distribution*

   $$\text{gamma}(\lambda \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

   is a conjugate prior for the exponential. Derive the parameter update given observations $x_1, \ldots, x_N$ and the prediction distribution $P(x_{N+1} \mid x_1, \ldots, x_N)$.

   (b) Show that the beta distribution is a conjugate prior for the *geometric distribution*

   $$P(X_i = k \mid \theta) = (1-\theta)^{k-1}\theta, \quad k = 1, 2, 3, \ldots$$

   which describes the number of time a coin is tossed until the first heads appears, when the probability of heads on each toss is θ. Derive the parameter update rule and prediction distribution.

   (c) Suppose $P(\boldsymbol{\theta} \mid \boldsymbol{\gamma})$ is a conjugate prior for the likelihood $P(\mathbf{X} \mid \boldsymbol{\theta})$; show that the *mixture prior*

   $$P(\boldsymbol{\theta} \mid \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_M) = \sum_{m=1}^{M} w_m P(\boldsymbol{\theta} \mid \boldsymbol{\gamma}_m)$$

   is also conjugate for the same likelihood, assuming the mixture weights w_m sum to 1. Derive an update equation similar to Equation (1).

   (d) Repeat part (c) for the case where the prior is a single distribution and the likelihood is a mixture, and the prior is conjugate for each mixture component of the likelihood.[2]

   (e) (Extra credit, 20) Explore the case where the likelihood is a mixture with fixed *components* and unknown *weights*; i.e., the weights are the parameters to be learned.

[1] The rate parameter λ for an exponential distribution is the inverse of the scale parameter b that we used in Assignment 4.

[2] Note that some priors can be conjugate for several different likelihoods; for example, the beta is conjugate for the Bernoulli and the geometric distributions and the gamma is conjugate for the exponential and for the gamma with fixed α.

2. *Bayesian Naive Bayes* (30)
   Now we will use some of the mathematical results on the spam problem.

   (a) Suppose that a Naive Bayes model has a Boolean output and continuous inputs modeled by exponential distributions. Describe how to perform Bayesian learning instead of maximum-likelihood learning.

   (b) Modify your code from Assignment 4 to use your new learning scheme. Choose appropriate priors and generate learning curves for the two methods (ML and Bayesian) on the spam problem. To make a meaningful comparison you may need to run several trials and average the results. Plot your results on one graph called `BNB.pdf`.

3. *Logistic regression for credit scoring* (40)
   Credit scoring is the process of determining if a credit or loan applicant is a good risk, based on data from an application form and previous credit history. Training data for predictive models consists of the original data for each applicant and the final outcome—in the simplest case, 1 for good behavior and 0 for defaulting on the loan. A somewhat simplified dataset, dealing with applicants for fixed-term loans at a bank, is available **here**.

   (a) Implement a data structure for a logistic regression model and a prediction method that gives the probability of outcome 1 for a given input example.

   (b) Implement a maximum-likelihood training method using *stochastic gradient descent*, which repeatedly picks a random example from the training set and takes a step in the gradient direction for that example. The gradient is easily derived from the log likelihood.

   (c) Generate a learning curve for the credit-scoring problem and plot your results on a graph called `LR.pdf`.

   (d) Describe briefly how you would make a decision about a loan applicant given your trained model.

   Turn in all your code, organized into clearly marked sections according to the parts of the assignment. Supply documentation and explanations where appropriate; describe any methods (cross-validation, multiple trials, etc.) you used to evaluate your methods and get good results.

   Submit your files collected together as `a5.tar.gz` using `submit a5` as described **here**