Homework Assignment 5–Due Dec 5

SDS 383C Statistical Modeling I

1 Document clustering with EM (10 points)

Recall the Bayesian model for documents.

Prior: $\pi \sim Dir(\alpha/K, \ldots, \alpha/K)$ For $k = 1 : K$ $\beta_k \sim Dir(\lambda, \ldots, \lambda)$
For each document $d = 1 : D$ draw $z_d \sim \pi$
For $i = 1 : N_d$ draw word $w_{di}|z_d = k \sim \beta_k$

1. (3 pts) Derive the E and M steps for obtaining MLE estimates of the parameters using Expectation Maximization.

2. Download the preprocessed BBCSport dataset from `http://mlg.ucd.ie/datasets/bbc.html`.

   (a) Using $\alpha = 0.3$, $\lambda = 1$ and $K = 5$, carry out the EM algorithm on this dataset. Note that this has 5 categories. We will use these labels for testing only.

      i. (1 pts) Plot the log likelihood and log posterior probabilities (you can lose the proportionality constants) as a function of iterations.

      ii. (3 pts) For $k = 1 : K$, list the 10 words (not word ids) with largest β_{ik}'s.

      iii. (3 pts) Using the categories from the dataset construct K β^* vectors of word proportions. For each category you can calculate the number of times a word appeared in documents in that category. This will give you the "true" proportions. How well do your learned β_k's line up with the β_k^*'s?

2 Bootstrap and Subsampling (10 pts)

In class we saw non-parametric bootstrap and subsampling. In this question we will work on some simulations to see cases where the bootstrap fails but subsampling works. In non-regular problems, the MLE is not asymptotically normal and the scaling constant is usually not $\sqrt{n}$. You have shown in an earlier homework that if $X_1, \ldots, X_n \sim U[0, \theta]$ then the MLE $\hat{\theta}$ satisfies $n(\theta - \hat{\theta}) \xrightarrow{d} Exp(\theta)$, where $Exp(\theta)$ is a exponential distribution with parameter θ.

(a) (3 pts) Simulate n datapoints from $U[0, 1]$. Now forget that you know $\theta = 1$. Use non-parametric bootstrap to compute the asymptotic variance of $\hat{\theta}$. For $n = 500 : 500 : 10000$ plot the relative error (averaged over 50 random runs) between the bootstrap estimate of variance and $1/n^2$.

(b) (2 pts) Use parametric bootstrap to compute the asymptotic variance of $\hat{\theta}$. For $n = 500 : 500 : 10000$ calculate the relative error between the parametric bootstrap estimate of variance and $1/n^2$. Plot the average over 50 random runs.

(c) (4 pts) Use subsampling to compute the asymptotic variance of $\hat{\theta}$. For $n = 500 : 500 : 10000$ plot the relative error between the subsampling estimate of variance and $1/n^2$. In the same figure, plot different curves for different values of b (averaged over 50 random runs). Be creative about your choice of different values of b. Remember that b has to grow with n but $b/n \to 0$. Don't forget to scale the subsampling estimates of variance properly.

(d) (1 pts) Compare the 3 methods.

3 Bootstrap theory

Let $X_1, \ldots, X_n \sim Bernoulli(p)$. and let $X_1^*, \ldots, X_n^*$ be a bootstrap sample. Let $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\hat{p}^* = \frac{1}{n}\sum_{i=1}^n X_i^*$.

1. (3 pts) What is the exact distribution of $n\hat{p}^*$, conditional on $X_1, \ldots, X_n$?
2. (3 pts) Find an explicit expression for $\text{var}(\hat{p}^*|X_1, \ldots, X_n)$.
3. (1 pt) What is the asymptotic distribution of $\sqrt{n}(\hat{p} - p)$?
4. (3 pts) What is the asymptotic distribution of $\sqrt{n}(\hat{p}^* - \hat{p})$, conditioned on $X_1, \ldots, X_n$?