CSC411 A5

Zhihong Wang 1002095207

2018-11

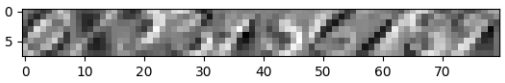
Q1:

(a).
Training set average conditional likelihood = -0.124624436669
Testing set average conditional likelihood = -0.196673203255

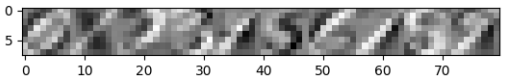
(b).
Training set accuracy = 0.9814285714285714
Testing set accuracy = 0.97275

(c).

Training eigenvectors:



Testing eigenvectors:



Q2:

[2pts] Categorial Distribution. Let's consider fitting the categorical distribution, which is a discrete distribution over K outcomes, which we'll number 1 through K. The probability of each category is explicitly represented with parameter θ_k. For it to be a valid probability distribution, we clearly need $\theta_k \geq 0$ and $\sum_k \theta_k = 1$. We'll represent each observation $\mathbf{x}$ as a 1-of-K encoding, i.e, a vector where one of the entries is 1 and the rest are 0. Under this model, the probability of an observation can be written in the following form:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{x_k}.$$

Denote the count for outcome k as N_k, and the total number of observations as N. In the previous assignment, you showed that the maximum likelihood estimate for the counts was:

$$\hat{\theta}_k = \frac{N_k}{N}.$$

Now let's derive the Bayesian parameter estimate.

(a).

[1pts] For the prior, we'll use the Dirichlet distribution, which is defined over the set of probability vectors (i.e. vectors that are nonnegative and whose entries sum to 1). Its PDF is as follows:

$$p(\boldsymbol{\theta}) \propto \theta_1^{a_1 - 1} \cdots \theta_K^{a_k - 1}.$$

A useful fact is that if $\boldsymbol{\theta} \sim \text{Dirichlet}(a_1, \ldots, a_K)$, then

$$\mathbb{E}[\theta_k] = \frac{a_k}{\sum_{k'} a_{k'}}.$$

Determine the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$, where $\mathcal{D}$ is the set of observations. From that, determine the posterior predictive probability that the next outcome will be k.

Determine Posterior Distribution:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{\int p(\theta')p(D|\theta')d\theta'}$$

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

$$\propto \prod_{i=1}^{n} \prod_{j=1}^{K} \theta_j^{\mathbb{1}(x_{ij}=1)} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1}$$

$$= \prod_{j=1}^{K} \theta_j^{N_j} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1} \text{ (Note: } N_j = \sum_{i=1}^{n} \mathbb{1}(x_{ij} = 1))$$

$$= \prod_{j=1}^{K} \theta_j^{N_j + \alpha_j - 1}$$

$$= \prod_{j=1}^{K} \theta_j^{\beta_j - 1} \text{ (Note: let } \beta_j = N_j + \alpha_j)$$

So, $p(\theta|D) \propto \theta_1^{\beta_1 - 1} ... \theta_K^{\beta_K - 1}$

Determine Posterior Predictive Probability:

Note: $x^{'} = k$ means $x_k^{'} = 1$

$$\theta_{pred} = Pr(x^{'} = k|D)$$

$$= \int p(\theta|D) Pr(x^{'} = k|\theta) d\theta_k$$

$$= \int \prod_{j=1}^{K} \theta_j^{\beta_j - 1} \theta_k d\theta_k$$

$$= E[\theta_k]$$

$$= \frac{\beta_k}{\sum_{k'} \beta_{k'}} \text{ (Note: } \beta_k = N_k + \alpha_k)$$

(b).
[1pt] Still assuming the Dirichlet prior distribution, determine the MAP estimate of the parameter vector $\boldsymbol{\theta}$. For this question, you may assume each $a_k > 1$.

Note: let $\beta_k = N_k + \alpha_k$

$$\hat{\theta}_{MAP} = argmax_\theta(p(\theta|D))$$

$$= argmax_\theta(\prod_{k=1}^{K} \theta_k^{\beta_k - 1})$$

$$= argmax_\theta(log \prod_{k=1}^{K} \theta_k^{\beta_k - 1})$$

$$= argmax_\theta(\sum_{k=1}^{K} log(\theta_k^{\beta_k - 1}))$$

$$= argmax_\theta(\sum_{k=1}^{K} (\beta_k - 1) log(\theta_k))$$

(Note: similar to A4 Q2(d). $\sum \theta_i = 1$)

Let $L(\theta_k) = \sum_{k=1}^{K} (\beta_k - 1) log(\theta_k) - \lambda(\sum_{k=1}^{K} \theta_k - 1)$

then $\frac{\partial L(\theta_k)}{\partial \theta_k} = \frac{\beta_k - 1}{\theta_k} - \lambda$

Let $\frac{\partial L(\theta_k)}{\partial \theta_k} = 0$

then $\frac{\beta_k - 1}{\theta_k} = \lambda$

then $\theta_k = \frac{\beta_k - 1}{\lambda}$

then $\sum \theta_k = \frac{1}{\lambda} \sum (\beta_k - 1)$

then $\sum_{k=1}^{K}(\beta_k - 1) = \lambda$, since $\sum \theta_i = 1$

then $\frac{\beta_k - 1}{\theta_k} = \sum_{k=1}^{K}(\beta_k - 1)$

Therefore $\hat{\theta}_{MAP} = \frac{\beta_k - 1}{\sum_{k=1}^{K}(\beta_k - 1)}$ (Note: $\beta_k = N_k + \alpha_k$)