Data Analysis

1. We have two finite sets of items $\mathcal{U} = \{u_1, \ldots, u_M\}$ and $\mathcal{V} = \{v_1, \ldots, v_N\}$. For example, U can be a set of adjectives and V be a set of nouns. Let $p(u_i, v_j)$ be the probability that the items u_i and v_j co-occur (for example, a pair of adjective and noun occurs as the subject of a sentence). Assume for some latent variable z taking values from $\{1, \ldots, K\}$, we have the conditional independence,

$$p(u_i, v_j|z = k) = p(u_i|z = k)p(v_j|z = k).$$

   (a) Let $P = [p(u_i, v_j)] \in R^{M\times N}$, the M-by-N matrix whose (i, j) element is $p(u_i, v_j)$; $U = [p(u_i|z = k)] \in R^{M\times K}$, the M-by-K matrix whose (i, k) element is $p(u_i|z = k)$; $V = [p(v_j|z = k)] \in R^{N\times K}$, the N-by-K matrix whose (j, k) element is $p(v_j|z = k)$; and the K-by-K diagonal matrix Σ with (k, k) element $p(z = k)$. Show that

$$P = U\Sigma V^T.$$

   (b) How does the above factorization of P differ from the singular value decomposition of P?

   (c) In an iid sample from $\{p(u_i, v_j)\}$, we observe the count data $\{n_{ij}, i = 1, \ldots, M, j = 1, \ldots, N\}$, i.e., items u_i and v_j co-occured n_{ij} items in the sample. Derive an EM algorithm that estimates the parameter matrices U, V, Σ based on those count data.

2. Consider a multiclass text classification problem where we have l labeled documents and u unlabeled ones where the label is missing $(x^{(1)}, y^{(1)}), \ldots, (x^{(l)}, y^{(l)}), x^{(l+1)}, \ldots, x^{(l+u)}$. All documents from class r were generated from a multinomial or naive Bayes distribution with parameter $\theta^{(r)}$.

   (a) Derive is the maximum likelihood estimator for $\theta^{(r)}$ using only the labeled data.

   (b) What is the mean squared error of this estimator?

   (c) We can construct an estimator for $\theta^{(1)}, \ldots, \theta^{(k)}$ that uses the unlabeled as well as the labeled data by maximizing the likelihood of the observed data

$$\sum_{i=1}^{l} \log p(x^{(i)}, y^{(i)}) + \sum_{i=l+1}^{l+u} \log p(x^{(i)}).$$

   Simplify the loglikelihood above as much as possible.

   (d) Show how the EM algorithm can be used to find the MLE in (c).

   In your answers please use V to denote the vocabulary, k to denote the number of classes, and $c(x^{(i)}, w)$ to denote the number of times word w appeard in document $x^{(i)}$.

3. Principal component analysis is a technique to embed high dimensional data $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$ in a low dimensionsal space $z^{(1)}, \ldots, z^{(n)} \in \mathbb{R}^l$ with $l \ll d$.

   (a) Write down a detailed description of the PCA algorithm. Specifically, explain how the high dimensional data are used to compute the dimensionality reduction and provide a formula for the coordinates of the reduced dimensional data $z^{(i)}$.

   (b) Describe a way to measure the amount of distortion caused by PCA and a principled way to determine what is an appropriate value of l.

   (c) Assume that $x^{(1)}, \ldots, x^{(n)} \sim N(0, \Sigma)$ where Σ is a diagonal matrix. What will be the PCA embedding in the limit of large data $n \to \infty$ in terms of Σ.

   (d) Repeat (c) above for a non-diagonal matrix Σ.

4. Suppose that you have a classification algorithm (an SVM, say) and a feature selection method which can work with it. Examples of such a feature selection method include forward subset selection and backward subset selection. If you have not heard of these, consider a program which randomly chooses subsets of the features, then performs SVM training to obtain a model using each subset to predict the target, and records the resulting training error for each subset; the final feature subset chosen is the one that yielded the best training error. Your colleague has performed this procedure on the dataset and has obtained a subset of features as a result.

   We would like to obtain the best value of the parameter C of the SVM. Your colleague would like to perform v-fold cross-validation, feeding it the new dataset having only the features selected by the feature selection program. Is this a good approach? If not, what is wrong with it, and what is a better approach? If so, justify the approach.