Homework 1 for #70240413
"Statistical Machine Learning"

Instructor: Prof. Jun Zhu

March 22, 2015

1 Mathematics Basics

Choose one problem from 1.1 and 1.2. A bonus will be rewarded if you finish both.

1.1 Calculus

The gamma function is defined by (assuming $x > 0$)

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du. \tag{1}$$

(1) Prove that $\Gamma(x+1) = x\Gamma(x)$.
(2) Also show that

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1} d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{2}$$

This result implies that the *Beta distribution* is normalized.

1.2 Optimization

Use the Lagrange multiplier method to solve the following problem:

$$\begin{aligned} \min_{x_1, x_2} \quad & x_1^2 + x_2^2 - 1 \\ s.t. \quad & x_1 + x_2 - 1 = 0 \\ & 2x_1 - x_2 \geq 0 \end{aligned} \tag{3}$$

Choose one problem from 1.3 and 1.4. A bonus will be rewarded if you finish both.

1.3 Probability

Suppose $p \sim \text{Beta}(p|\alpha, \beta)$ and $x|p \sim \text{Bernoulli}(x|p)$. Show that $p|x \sim \text{Beta}(p|\alpha + x, \beta + 1 - x)$, which implies that the Beta distribution can serve as a conjugate prior to the Bernoulli distribution.

1.4 Stochastic Process

We toss a fair coin for a number of times and use H(head) and T(tail) to denote the two sides of the coin. Please compute the expected number of tosses we need to observe a first time occurrence of the following consecutive pattern

$$H, \underbrace{T, T, \cdots, T}_{k}.$$

2 Mixture of Multinomials

2.1 MLE for multinomial

Derive the maximum-likelihood estimate for the parameter $\boldsymbol{\mu} = (\mu_i)_{i=1}^d$ of a multinomial distribution:

$$P(\boldsymbol{x}|\boldsymbol{\mu}) = \frac{n!}{\prod_i x_i!} \prod_{i=1}^{d} \mu_i^{x_i} \tag{4}$$

where $x_i \in \mathbb{N}$, $\sum_i x_i = n$ and $0 < \mu_i < 1$, $\sum_i \mu_i = 1$.

2.2 EM for mixture of multinomials

Consider the following mixture-of-multinomials model to analyze a corpus of documents that are represented in the bag-of-words model.

Specifically, assume we have a corpus of D documents and a vocabulary of W words from which every word in the corpus is taken. We are interested in counting how many times each word appears in each document, regardless of their positions and orderings. We denote by $T \in \mathbb{N}^{D \times W}$ the word occurrence matrix where the wth word appears T_{dw} times in the dth document.

According to the mixture-of-multinomials model, each document is generated i.i.d. as follows. We first choose for each document d a *latent* "topic" c_d (analogous to choosing for each data point a component z_n in the mixture-of-Gaussians) with

$$P(c_d = k) = \pi_k,\ k = 1, 2, \cdots, K; \tag{5}$$

And then given this "topic" $\boldsymbol{\mu}_k = (\mu_{1k}, \ldots, \mu_{Wk})$ which now simply represents a categorical distribution over the entire vocabulary, we generate the word bag of the document from the corresponding multinomial distribution[1]

$$P(d|c_d = k) = \frac{n_d!}{\prod_w T_{dw}!} \prod_w \mu_{wk}^{T_{dw}},\ \text{where } n_d = \sum_w T_{dw}. \tag{6}$$

[1] Make sure you understand the difference between a categorical distribution and a multinomial distribution. You may think about a Bernoulli distribution and a binomial distribution for reference.

Hence in summary

$$P(d) = \sum_{k=1}^{K} P(d|c_d = k)P(c_d = k) = \frac{n_d!}{\prod_w T_{dw}!} \sum_{k=1}^{K} \pi_k \prod_w \mu_{wk}^{T_{dw}}. \tag{7}$$

Given the corpus T, please design and implement an EM algorithm to learn the parameters $\{\boldsymbol{\pi}, \boldsymbol{\mu}\}$ of this mixture model and test it on the "20newsgroup" dataset[2].

Set the number of topics K to be $5, 10, 20, 30$ respectively and show the most-frequent words in each topic for each case. Observe the result and try to find the "best" K value for this dataset and explain why.

3 PCA

3.1 Minimum Error Formulation

Complete the proof on the lecture slide which shows that PCA can be equivalently formulated as minimizing the mean-squared-error of a low-dimensional approximation from a subset of orthonormal basis.

3.2 MNIST

Implement the PCA algorithm and test it on the MNIST dataset[3]. Rerun your PCA implementation without centering the dataset (substracting the sample mean) and report the difference between the results.

[2] `http://ml.cs.tsinghua.edu.cn/~wenbo/data/20newsgroup_train.zip`
[3] `http://www.cs.nyu.edu/~roweis/data.html`