

# Homework 4

Introduction to Machine Learning  
Fall 2018  
Instructor: Anna Choromanska

Homework is due 11/02/2018.

## Problem 1 (20 points): EM

Consider a random variable  $x$  that is categorical with  $M$  possible values  $1, 2, \dots, M$ . Suppose  $x$  is represented as a vector such that  $x(i) = 1$  if  $x$  takes the  $i$ th value, and  $\sum_{i=1}^M x(i) = 1$ . The distribution of  $x$  is described by a mixture of  $K$  discrete multinomial distributions such that:

$$p(x) = \sum_{k=1}^K \pi_k p(x|\mu_k)$$

and

$$p(x|\mu_k) = \prod_{j=1}^M \mu_k(j)^{x(j)},$$

where  $\pi_k$  denotes the mixing coefficient for the  $k$ th component (aka the prior probability that the hidden variable  $z = k$ ), and  $\mu_k$  specifies the parameters of the  $k^{\text{th}}$  component. Specifically,  $\mu_k(j)$  represents the probability  $p(x(j) = 1|z = k)$ , and  $\sum_j \mu_k(j) = 1$ . Given an observed data set  $\{x_i\}, i = 1, 2, \dots, N$ , derive the  $E$  and  $M$  step of the EM algorithm for optimizing the mixing coefficients and the component parameters  $\mu_k(j)$  for this distribution (below we provide the generic formula for the E and M steps, where  $\theta$  denotes all the parameters of the mixture model).

- E-step (5 points): For each  $i$ , calculate  $Q_i(z_i) = p(z_i|x_i;\theta)$ , i.e. the probability that observation  $i$  belongs to each of the  $K$  clusters.
- M-step (15 points): Set

$$\theta := \arg \max_{\theta} \sum_{i=1}^N \sum_{z_i} Q_i(z_i) \log \frac{p(x_i, z_i; \theta)}{Q_i(z_i)}.$$

## Problem 2 (10 points): Clustering

**Lemma 1** *Let  $\phi(W_1)$  be the optimum value of the  $k$ -means objective for the  $k$ -clustering of data set  $W_1$ , and let  $\phi(W_2)$  be the optimum value of the  $k$ -means objective for the  $k$ -clustering of data set  $W_2$ . Finally, let  $\phi(W_1 \cup W_2)$  be the optimum value of the  $k$ -means objective for the  $k$ -clustering of data set  $W_1 \cup W_2$ . Prove that*

$$\phi(W_1) + \phi(W_2) \leq \phi(W_1 \cup W_2).$$

## Problem 3 (20 points): MLE and MAP

Consider a univariate normal distribution. Given  $N$  one-dimensional scalar samples,  $\{x_1, \dots, x_N\}$  and  $x_i \in \mathbb{R}$ , independently drawn from a normal distribution with KNOWN variance  $\sigma^2$  and UNKNOWN mean  $\mu$ , derive

- a) [8 points] Maximum Likelihood Estimator for the mean  $\mu$ .
- b) [8 points] Maximum a Posteriori Estimator for the mean  $\mu$ . Assume that the prior distribution for the mean is a normal distribution with mean  $\nu$  and variance  $\beta^2$ .
- c) [4 points] How do the estimators behave when the number of samples  $N$  goes to infinity?