Homework 0: Statistical Machine Learning

UC Irvine CS274B: Learning in Graphical Models

REVIEW: NOT COLLECTED OR GRADED

Question 1:

This question asks you to devise maximum likelihood and Bayesian estimators for a simple model of an uncalibrated sensor. Let the sensor output, X, be a random variable that ranges over the positive real numbers. We assume that, when tested over a range of environments, its outputs are uniformly distributed on some unknown interval $[0, \theta]$, so that

$$p(x \mid \theta) = \begin{cases} 1/\theta & \text{if } 0 \le x \le \theta, \\ 0 & \text{otherwise,} \end{cases}$$
$$= \frac{1}{\theta} \mathbb{I}_{0,\theta}(x) \, .$$

Here, $\mathbb{I}_{0,\theta}(x)$ denotes an *indicator function* which equals 1 when $0 \le x \le \theta$, and 0 otherwise. We denote this distribution by $X \sim \text{Unif}(0, \theta)$. To characterize the sensor's sensitivity, we would like to infer θ.

a) Given N i.i.d. observations $x = \{x_1, \ldots, x_N\}$, $X_n \sim \text{Unif}(0, \theta)$, what is the likelihood function $p(x \mid \theta)$? What is the maximum likelihood (ML) estimator for θ? Give an informal proof that your estimator is in fact the ML estimator.

b) Suppose that we place the following prior distribution on θ:

$$p(\theta) = \alpha \beta^{\alpha} \theta^{-\alpha-1} \mathbb{I}_{\beta,\infty}(\theta) \, .$$

This is known as a Pareto distribution. *We denote it by $\theta \sim \text{Pareto}(\alpha, \beta)$. Plot the three prior probability densities corresponding to the following three hyperparameter choices: $(\alpha, \beta) = (0.1, 0.1)$; $(\alpha, \beta) = (2.0, 0.1)$; $(\alpha, \beta) = (1.0, 2.0)$. Briefly describe the influence these parameters have on the properties of the Pareto distribution.*

c) If $\theta \sim \text{Pareto}(\alpha, \beta)$ and we observe N uniformly distributed observations $X_n \sim \text{Unif}(0, \theta)$, derive the posterior distribution $p(\theta \mid x)$. Is this a member of any standard family?

d) For the posterior derived in part (c), what is the corresponding MAP estimator of θ? How does this compare to the ML estimator?

e) Recall that the quadratic loss is defined as $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. For the posterior derived in part (c), what estimator of θ minimizes the posterior expected quadratic loss? Simplify your answer as much as possible.

f) Suppose that we observe three observations $x = (0.7, 1.3, 1.7)$. Determine the posterior distribution of θ for each of the priors in part (b), and plot the corresponding posterior densities. What is the MAP estimate for each hyperparameter choice? What estimate minimizes the quadratic loss for each hyperparameter choice?

Question 2:

We now consider a binary categorization problem, where $y_n \in \{0, 1\}$ is the output label for example n, and $x_n \in \mathbb{R}^2$ is a two-dimensional vector of input features. Assume that the two classes are equally likely *a priori*, so that $p(y_n) = \text{Ber}(y_n \mid 0.5)$. Under the true data generation process, the features are distributed according to class-specific Gaussians:

$$p(x_n \mid y_n = 1) = \text{Normal}(x_n \mid \mu_1, \Sigma), \qquad p(x_n \mid y_n = 0) = \text{Normal}(x_n \mid \mu_0, \Sigma).$$

The mean vectors μ_1, μ_0 are discussed below. The shared covariance matrix equals:

$$\Sigma = \frac{4}{3}\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix} = \begin{bmatrix} 4/3 & 2/3 \\ 2/3 & 4/3 \end{bmatrix}, \qquad \Sigma^{-1} = \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix}.$$

a) Suppose that $\mu_0 = [0, 0]^T$, $\mu_1 = [2, 0]^T$. Given knowledge of the true joint distribution $p(x_n, y_n)$, derive a classification rule $\hat{y}(x_n)$ that minimizes the probability of error. Plot the corresponding decision boundary graphically.

b) Suppose that $\mu_0 = [0, 0]^T$, $\mu_1 = [2, 2]^T$. Given knowledge of the true joint distribution $p(x_n, y_n)$, derive a classification rule $\hat{y}(x_n)$ that minimizes the probability of error. Plot the corresponding decision boundary graphically.

Now suppose that we do not have knowledge of the true data generating process, but instead assume a naive Bayes model with Gaussian features:

$$p(x_n \mid y_n = 1) = \text{Normal}(x_{n1} \mid \theta_{11}, \nu_{11})\text{Normal}(x_{n2} \mid \theta_{12}, \nu_{12}),$$
$$p(x_n \mid y_n = 0) = \text{Normal}(x_{n1} \mid \theta_{01}, \nu_{01})\text{Normal}(x_{n2} \mid \theta_{02}, \nu_{02}).$$

Consider a training dataset with N observations (x_n, y_n) independently sampled from the true joint distribution $p(x, y)$. In each question below, assume that the parameters θ and ν of the naive Bayes model are estimated via the maximum likelihood (ML) criterion.

c) Suppose that $\mu_0 = [0, 0]^T$, $\mu_1 = [2, 0]^T$. As $N \to \infty$, what classification rule will the naive Bayes classifier approach? Will it be as accurate as the optimal rule from part (a)? Justify your answer and plot the corresponding decision boundary graphically.

d) Suppose that $\mu_0 = [0, 0]^T$, $\mu_1 = [2, 2]^T$. As $N \to \infty$, what classification rule will the naive Bayes classifier approach? Will it be as accurate as the optimal rule from part (b)? Justify your answer and plot the corresponding decision boundary graphically.

Question 3:

In this question, you will derive an *expectation maximization* (EM) algorithm for clustering vectors of counts. Let $x = \{x_1, \ldots, x_N\}$ denote the N input observations. Each observation is a vector $x_n = [x_{n1}, x_{n2}, \ldots, x_{nD}]$ of D non-negative integers, $x_{nj} \in \{0, 1, 2, \ldots\}$. For example, to find clusters of users with similar web browsing patterns, we could let x_{nj} be the number of times user n visits website j in a given month.

We model this data via a mixture model with K components, parameterized as follows:

$$p(x_n \mid \pi, \theta) = \sum_{k=1}^{K} \pi_k \left[\prod_{j=1}^{D} \text{Geom}(x_{nj} \mid \theta_{kj}) \right],$$
$$\text{Geom}(x \mid \theta) = (1-\theta)^x \theta \quad \text{for} \quad x \in \{0, 1, 2, \ldots\}.$$

The parameters of the model are $\{\pi, \theta\}$, where $0 \leq \pi_k \leq 1$ is the probability of choosing mixture component k, and $0 < \theta_{kj} \leq 1$ is the success probability for the geometric distribution modeling the count x_{nj} for an example drawn from cluster k. Note that there are KD of these geometric parameters, which define the mean of each count within each cluster. To denote the unobserved cluster assignments, let $z_{nk} = 1$ if mixture component k generates observation x_n, and $z_{nk} = 0$ otherwise.

Hint: Remember that the following constrained optimization problem, involving a K-dimensional vector of probabilities and constants $C_k \geq 0$, has a simple solution:

$$\text{If } \hat{\pi} = \arg\max_{\pi} \sum_{k=1}^{K} C_k \log \pi_k, \text{ subject to } \sum_{k=1}^{K} \pi_k = 1, \text{ then } \hat{\pi}_k = \left[\sum_{k'=1}^{K} C_{k'} \right]^{-1} C_k. \quad (1)$$

a) *Suppose first that we have "complete" data, so that the mixture components z_n associated with each observation are directly observed. Give an expression for the complete-data log-likelihood $\log p(x, z \mid \pi, \theta)$. Derive maximum likelihood (ML) estimators for $\{\pi, \theta\}$ given (x, z). Use N_k to denote the number of examples of cluster k.*

b) *Now consider the E-step of EM, where mixture assignments z_n are unobserved. Derive a formula for the conditional probability distributions $p(z_n \mid x, \pi, \theta)$ of cluster assignments given model parameters $\{\pi, \theta\}$ from the last M-step. Denote the parameters of these distributions by $r_{nk} = P(z_{nk} = 1 \mid x, \pi, \theta) = \mathbb{E}[z_{nk} \mid x, \pi, \theta]$.*

c) *Give an expression for the expected complete-data log-likelihood $\mathbb{E}[\log p(x, z \mid \pi, \theta)]$, where the expectation is with respect to some distribution $q(z)$ determined in the E-step. Derive formulas for the M-step of the EM algorithm for estimating $\{\pi, \theta\}$. Your formulas should be explicit functions of the expected values r_{nk} from part (b), and the observed data x.*

d) *How does the computational complexity of the E-step and M-step scale with the number of observations N, the dimension D, and the number of clusters K? Justify your answer.*

e) *Suppose that many of the counts in your dataset are exactly zero. You store this data as a sparse matrix, represented by a list of the locations and values of the $L < ND$ positive counts (the (n, j) pairs where $x_{nj} > 0$). By exploiting this sparsity, how much could you reduce the computational cost of the E-step and M-step? Justify your answer.*

Question 4:

In this final question, we examine the probabilistic principal component analysis (PPCA) model. To simplify things, we make the following assumptions:

- The N training vectors $x_n \in \mathbb{R}^{D\times 1}$, $n = 1, \ldots, N$, have already been centered, so that $\sum_{n=1}^{N} x_n = 0$. We thus constrain the PPCA model to also have zero mean.
- The desired latent space is one-dimensional, so that observations are represented by coordinates $z_n \in \mathbb{R}$.

In this case, the PPCA generative model can be written as follows:

$$p(z_n) = \text{Normal}(z_n \mid 0, 1), \qquad p(x_n \mid z_n, w, \lambda) = \text{Normal}(x_n \mid wz_n, \lambda I_D).$$

Here I_D is a $D \times D$ identity matrix, and $w \in \mathbb{R}^{D\times 1}$ and $\lambda > 0$ are parameters to be estimated from the N training observations $x = \{x_1, x_2, \ldots, x_N\}$. We will estimate these parameters via the EM algorithm.

a) Suppose that the PPCA model parameters w, λ are known, and consider the posterior distribution $p(z_n \mid x_n, w, \lambda)$. Computation of this distribution is the E-step of the EM algorithm for PPCA. What standard family is it a member of? Give explicit formulas for all parameters of the posterior distribution, in terms of x_n, w, and λ.

b) Give an expression for the expected complete-data log-likelihood $\mathbb{E}[\log p(x, z \mid w, \lambda)]$, where the expectation is with respect to a distribution on z in the family determined in part (a). What particular moments of z_n must be computed to explicitly evaluate this expression?

c) Take the partial derivative of the expected log-likelihood from part (b) with respect to λ, set to zero, and simplify to determine the M-step estimate $\hat{\lambda}$ of the variance parameter.

d) Take the partial derivative of the expected log-likelihood from part (b) with respect to w_k, an element of the principal subspace vector w. Set this expression to zero, and simplify to determine the M-step estimate $\hat{w}$ of the principal subspace.

e) Suppose that the EM algorithm converges to a particular set of parameters $\hat{w}, \hat{\lambda}$. Are these ML parameter estimates unique? If so, provide an argument for why this is the case. If not, construct an alternative set of parameters $\bar{w}, \bar{\lambda}$ which have equal log-likelihood, i.e. which satisfy $\log p(x \mid \hat{w}, \hat{\lambda}) = \log p(x \mid \bar{w}, \bar{\lambda})$.