CSE Ph.D. Qualifying Exam, Spring 2024

This is a closed-book, closed-notes exam.

Data Analysis

Please answer three of the following four questions. All questions are graded on a scale of 10. If you answer all four, all answers will be graded and the three lowest scores will be used in computing your total. Show all your work and write in a readable way.

1. **Probabilistic PCA**
   The formulation of PCA was based on a linear projection of the data onto a subspace of lower dimensionality than the original data space. It can be shown that PCA can also be expressed as the maximum likelihood solution of a probabilistic latent variable model. This reformulation of PCA, known as **probabilistic PCA (PPCA)**. PPCA is a simple example of the linear-Gaussian framework, in which all of the marginal and conditional distributions are Gaussian. We can formulate PPCA by first introducing an explicit latent variable $z \in \mathbb{R}^{M\times 1}$ corresponding to the principal-component subspace. Next we define a Gaussian prior distribution $p(z)$ over the latent variable, together with a Gaussian conditional distribution $p(x|z)$ for the observed variable $x \in \mathbb{R}^{D\times 1}$ conditioned on the value of the latent variable. Specifically, the prior distribution over z is given by a zero-mean unit-covariance Gaussian $p(z) = \mathcal{N}(z|0, \mathbf{I})$. Similarly, the conditional distribution of the observed variable x, conditioned on the value of the latent variable z, is again Gaussian, of the form $p(x|z) = \mathcal{N}(x|\mathbf{W}z + \mu, \sigma^2\mathbf{I})$ in which the mean of x is a general linear function of z governed by the $D \times M$ matrix $\mathbf{W}$ and the D-dimensional vector μ. All μ, $\mathbf{W}$ and σ^2 are unknown parameters.

   a. [1.5 points] Derive the marginal distribution $p(x)$ with μ, $\mathbf{W}$ and σ^2.
   b. [1.5 points] Suppose we replace the zero-mean, unit-covariance latent space distribution $p(z)$ in the PPCA model by a general Gaussian distribution of the form $\mathcal{N}(z|m, \Sigma)$. By redefining the parameters of the model, show that this leads to an identical model for the marginal distribution $p(x)$ over the observed variables for any valid choice of m and Σ.
   c. [1.5 points] Note that $p(x|z)$ factorizes with respect to the elements of x, in other words, this is an example of the naive Bayes model. Draw a directed probabilistic graph for the PPCA model and naive Bayes to show why.
   d. [5.5 points] Maximum likelihood PCA: We next consider the determination of the model parameters using maximum likelihood. Given a data set $\mathbf{X} = \{x_n\}_{n=1}^N$ of observed data points, where $x_n \in \mathbb{R}^{D\times 1}$,

   - d.1 [1 points] The corresponding log likelihood function is given by

     $$\ln p(\mathbf{X}|\mu, \mathbf{W}, \sigma^2) =$$

   - d.2 [1 points] Setting the derivative of the log likelihood with respect to μ equal to zero gives the expected result

     $$\mu =$$

- d.3 [1.5 points] Back-substituting the optimal μ to the log likelihood function, we can then write the log likelihood function in the form

$$\ln p(\mathbf{X}|\mathbf{W}, \sigma^2) =$$

- d.4 [2 points] Derive the closed-form $\mathbf{W}$ from the above log likelihood function as a function of σ^2 and data $\mathbf{X}$,

$$\mathbf{W} =$$

2. **Maximum Likelihood and Maximum A Posteriori Estimations**

Assume you are helping GaTech to develop an on-campus test for COVID-19. Your test has a false positive rate of α and a false negative rate of β.

(a) [1 pt] Assume that COVID-19 is evenly distributed through the population and that the prevalence of the disease is γ. What is the accuracy of your test on the general population?

(b) [1 pt] Assume there are n people on campus all of whom they know have COVID. What is the likelihood that the test makes n_+ correct predictions?

(c) [4 pts] Derive the maximum likelihood estimate for β. You may assume all other parameters are fixed.

(d) [4 pts] Derive the Maximum A Posteriori (MAP) estimate for β assuming it has a prior $P(\beta) = \text{Beta}(a, b)$. You may assume all other parameters are fixed. *Hint:* the probability density function of $\text{Beta}(a, b)$ is $p(x; a, b) = Z \cdot x^{a-1}(1-x)^{\beta-1}$ with Z as a constant.

3. **Neural Networks**

a. [1 point] A perceptron is an algorithm for learning a binary classifier that can be described by the following learning rule:

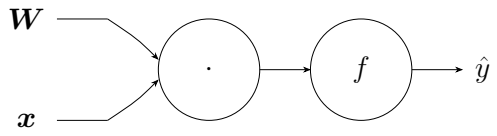
$$y = \begin{cases} 0 & \text{if } \boldsymbol{w} \cdot \boldsymbol{x} + b \leq 0 \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

where $\boldsymbol{w}$ are the weights, $\boldsymbol{x}$ is the input vector and b is the bias. Explain why a single perceptron can compute the logical AND and OR functions easily, but it cannot compute the logical XOR.

b. [3 points] Design a feed-forward neural network to solve the XOR problem. The network should have a single hidden layer of two neurons and an output layer of a single neuron. Use the ReLU activation function: $\text{ReLU}(x) = \max(0, x)$. Show your calculations for every possible input.

c. [3 points] The following figure shows the computational graph of a simple neural network, $\hat{y} = f(\boldsymbol{x}, \boldsymbol{W}) = \|\boldsymbol{W} \cdot \boldsymbol{x}\|^2$, where $\boldsymbol{x} \in \mathbb{R}^n$ is the input vector, $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ is the weights matrix of the network and $f(\boldsymbol{a}) = \|\boldsymbol{a}\|^2$. Note that x_i refers to the i-th

element of the vector $\boldsymbol{x}$ and W_{ij} refers to the element at the i-th row and j-th column of the matrix $\boldsymbol{W}$.



Let $\boldsymbol{q} = \boldsymbol{W} \cdot \boldsymbol{x}$, show the following

$$\frac{\partial f}{\partial q_i} = 2q_i; \qquad \frac{\partial f}{\partial W_{ij}} = 2q_i x_j; \qquad \frac{\partial f}{\partial x_i} = \sum_k 2q_k W_{k,i},$$

and give their vectorized forms respectively.

d. [2 points] Given the following values of $\boldsymbol{W}$ and $\boldsymbol{x}$, calculate the network estimate $\hat{y}$ by doing the forward computation once. Let the ground-truth label y be 0 and the loss function be $\mathcal{L}(\hat{y}, y) = |\hat{y} - y|$, update the weights matrix once using the gradient descent rule: $\boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \eta \nabla_{\boldsymbol{W}} \mathcal{L}$, where $\eta = 1$ is the learning rate.

e. [1 point] When training a neural network, why do we want to exclude regularization from the bias terms?

4. **Generative Models**

Consider Alice and Bob are asked to fit a generative model $p_\theta(x)$ over a given dataset $\mathcal{D} = \{x_i\}_{i=1}^n$, where θ denotes the parameters of the model.

(1) Alice exploited the *latent variable model* for $p_\theta(x) := \int p_\alpha(x|z)p_\beta(z)dz$ with $\theta = \{\alpha, \beta\}$.

i) [3 points] Please derive the evidence lower bound (ELBO) for the latent variable model, *i.e.*,

$$\log p_\theta(x) \geq \mathbb{E}_{z\sim q(z|x)} \left[\log p_\alpha(x|z)\right] - KL(q(z|x)||p_\beta(z)). \tag{2}$$

(2) Bob used the *energy-based model (EBM)* for $p_\theta(x)x := \frac{1}{Z_\theta} \exp\big(f_\theta(x)\big)$ with $Z_\theta := \int \exp\big(f_\theta(x)\big)dx$.

i) [3 points] Please derive the gradient of the MLE of EBM, *i.e.*,

$$\nabla_\theta \widehat{\mathbb{E}}_{x\sim\mathcal{D}} \left[\log p_\theta(x)\right] = \widehat{\mathbb{E}}_{x\sim\mathcal{D}} \left[\nabla_\theta f_\theta(x)\right] - \mathbb{E}_{x\sim p_\theta(x)} \left[\nabla_\theta f_\theta(x)\right]. \tag{3}$$

ii) [4 points] The $\mathbb{E}_{x\sim p_\theta(x)} \left[\nabla_\theta f_\theta(x)\right]$ is intractable, which makes the gradient (3) difficult to calculate, and thus, the learning of EBM. If $x \in \mathbb{R}^d$ is continuous, please design an approximation for $\mathbb{E}_{x\sim p_\theta(x)} \left[\nabla_\theta f_\theta(x)\right]$ in (3).