CSE Ph.D. Qualifying Exam, Spring 2025: Data Analysis

Instructions:

- This is a CLOSED BOOK exam. No books or notes are allowed.
- 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.
- Please write clearly and concisely, explain your reasoning, and show all work. Points will be awarded for clarity as well as correctness.
- Good luck!

1. **Q1: Neural Networks** Consider a variational autoencoder (VAE) with the following setup:

   The approximate posterior $q_\phi(z|x)$ is modeled as a Gaussian distribution:

   $$q_\phi(z|x) = \mathcal{N}\left(z; \mu_\phi(x), \sigma^2_\phi(x)\right), \tag{1}$$

   where $\mu_\phi(x)$ and $\sigma^2_\phi(x)$ are parameterized by the encoder neural network. The evidence lower bound (ELBO) objective is

   $$\mathcal{L}\left(\theta, \phi; x\right) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - KL\left(q_\phi(z|x)\|p(z)\right). \tag{2}$$

   (1) [1 points] Explain why the sampling process $z \sim q_\phi(z|x)$ poses a challenge for backpropagation. Use mathematical expressions to support your explanations.

   (2) [1 points] Describe how the reparameterization trick is used to address this challenge and explain why it is important to isolate the randomness during the reparameterization. Use mathematical expressions to support your explanations.

   (3) [2 points] Suggest a workaround if the reparameterization trick were not used.

   (4) [3 points] Given a weights matrix $\boldsymbol{W}$, an input vector $\boldsymbol{x}$, the sigmoid activation function $\sigma(z) = 1/\left(1 + \exp\left(-z\right)\right)$, we can construct a simple neural network $f(\boldsymbol{x}; \boldsymbol{W}) = \sigma(\boldsymbol{W}^\intercal \boldsymbol{x})$. Let the scalar model output be $\hat{y}$. Write down an expression for $\nabla_{\boldsymbol{W}}\hat{y}$.

   (5) [3 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 $\mathcal{L}(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$, update the weights matrix once using the gradient descent rule: $\boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \eta\nabla_{\boldsymbol{W}}\mathcal{L}$, where $\eta = 4$ is the learning rate.

2. **Q2: Multiclass Classification.** We consider the multiclass classification with softmax logistic regression. Specifically, we consider the k-class softmax parametrized conditional model

   $$p(C_k = 1|x) = \frac{\exp(\phi_k(x))}{\sum_{j=1}^k \exp(\phi_j(x))}, \tag{3}$$

   where $\phi_i(x)$ can be a linear model, i.e., $\phi_i(x) = w_i^\top x$, or a 2-layer perception neural network, i.e., $\phi_i(x) = w_i^\top \sigma(W^\top x)$ with $\sigma(\cdot)$ is a differentiable nonlinear activation function (also known as "neuron"), e.g., $\texttt{relu}(\cdot) = \max(0, \cdot)$ or $\texttt{tanh}(\cdot)$. Given dataset $\{x_i, y_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$, $y_i = [y_i^j]_{j=1}^k$ as a k-dim binary vector, and $y_i^j \in \{0, 1\}$ as a binary variable, for all $i = 1, \ldots, n$.

   (1) [4 points] Denote the parameters in $\phi\cdot$ as V ($V = \{w_i\}_{i=1}^k$ in linear model and $V = [\{w_i\}_{i=1}^k, W]$ in 2-layer perception neural network), please provide the maximum likelihood (MLE) of (3), $\ell(V)$, upon the given data.

(2) [3 points] Could you please calculate the gradient of MLE w.r.t. w_i in linear model?

(3) [3 points] Could you please calculate the gradient of MLE w.r.t. W in 2-layer perception neural network? (hint: use chain-rule for backpropagation.)

3. **Q3:** Given a dataset $\{(x^i, y^i)\}_{i=1,2,\cdots,n}$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{-1, 1\}$ for all $i \in \{1, 2, \cdots, n\}$, please answer the following questions about support vector machine (SVM).

The primal problem of SVM is derived from a margin-based minimization problem:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}|w|^2 + C\sum_{i=1}^{n}\xi^i && (4) \\ \text{s.t.} \quad & y^i(w^\top x^i + b) \geq 1 - \xi^i \quad \forall i && (5) \\ & \xi^i \geq 0 && (6) \end{aligned}$$

The dual problem is given by:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y^i y^j ({x^i}^\top x^j) && (7) \\ \text{s.t.} \quad & C \geq \alpha_i \geq 0 \quad \forall i && (8) \\ & \sum_{i=1}^{n}\alpha_i y^i = 0 && (9) \end{aligned}$$

(1) [3 points] Let us modify the penalty in the primal objective from ξ^i to $|\xi^i|^2$. The primal problem becomes:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}|w|^2 + \frac{1}{2}C\sum_{i=1}^{n}|\xi^i|^2 && (10) \\ \text{s.t.} \quad & y^i(w^\top x^i + b) = 1 - \xi^i \quad \forall i && (11) \end{aligned}$$

This is known as **least-squares support vector machines**. Please show how to derive the solution to the least squares support vector machines.

(2) [2 points] Please show how to make inference based on your solution in Q3-1. Please compare this with the inference of SVM.

(3) [2 points] What is the computation cost of solving the least-squares support vector machine? Please compare this with the computation cost of SVM.

(4) [3 points] Please show how to apply the kernel trick to the least squares SVM. Assume you have access to a given kernel $K(x, x') = \langle\phi(x), \phi(x')\rangle$, but you solution should only include the kernel function K but not the function mapping ϕ. Please also show how to make inference of kernel least-squares SVM.

4. **Q4: Expectation-Maximization.**

Derive the EM algorithm for a Gaussian Mixture Model (GMM). Assume you are given a dataset $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$, where each x_i is independently drawn from one of K Gaussian distributions. Each distribution k is characterized by parameters including a mean μ_k, covariance Σ_k, and a mixing coefficient π_k.

The complete data likelihood for the GMM includes the observed data $\mathbf{X}$ and latent variables $\mathbf{Z}$, where latent variable $z_{ik} = 1$ if x_i is generated by the k-th Gaussain and 0 otherwise.

[1] (2 points) Derive the likelihood of the observed data $\mathbf{X}$ along with the complete likelihood of the data involving the latent variables $\mathbf{Z}$

$$L(\theta; \mathbf{X}, \mathbf{Z}) = \tag{12}$$

An expectation-maximization algorithm includes E-step: compute the expected value of the complete data log-likelihood; and M-step: update the parameters.

[2] (2 points) Describe the E-step computation for updating the expected values of the latent variables $\mathbf{Z}$, and derive the responsibilities $\gamma(z_{ik})$, that indicates the probability that x_i belongs to the k-th Gaussian.

$$\gamma(z_{ik}) = \tag{13}$$

[3] (3 points) Describe the parameter updates in the M-step based on the responsibilities $\gamma(z_{ik})$ calculated in the E-step.

$$\pi_k^{(\text{new})} = \tag{14}$$

$$\mu_k^{(\text{new})} = \tag{15}$$

$$\Sigma_k^{(\text{new})} = \tag{16}$$

[4] (3 points) Discuss how to determine the convergence of the EM algorithm in practice. Does it guarantee to obtain global minima? How to make a good choice of initialization?