Mathematical foundations for ML - UM6P class

Final Exam

February 7th, 2025

The goal of the exam is to consolidate the concepts we have seen in class. Some rules:

- **Each exercise is worth 5 points, with Exercise 1 worth 10 points.**
- **You don't need to solve all the exercises to get the maximal score.**
- If you have a partial answer to a question, write it down: it might get you a partial grade.
- Be succinct, but as mathematically thorough as you can.

MCQ

Exercise 1.
There is exactly one correct answer per question - YOUR ANSWER TO THIS EXERCISE SHOULD BE EXACTLY ONE STRING OF 10 LETTERS. EXAMPLE: ABDACCADDB. Remember that a random answer would get you more points (in expectation) than no answer.

1. What is the covariance between two random variables X and Y?
   (a) $\mathbb{E}[XY]$
   (b) $\mathbb{E}[X]\mathbb{E}[Y]$
   (c) $\mathbb{E}[XY] + \mathbb{E}[X]\mathbb{E}[Y]$
   (d) $\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$
2. For a linear regression problem with input matrix $\mathbf{X}$ and target vector $\mathbf{y}$, what is the Maximum Likelihood Estimator (MLE) for the weights $\mathbf{w}$?
   (a) $\mathbf{X}^\top \mathbf{X}\mathbf{y}$
   (b) $(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y}$
   (c) $\mathbf{X}^\top \mathbf{y}$
   (d) $(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$
3. If X and Y are independent random variables, which statement is true?
   (a) $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
   (b) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - \mathrm{Var}(X - Y)$
   (c) $\mathbb{E}[XY^2] = \mathbb{E}[X^2 Y]$
   (d) None of the above
4. If $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$, which statement is true?
   (a) X and Y are independent
   (b) X and Y are correlated
   (c) X and Y are uncorrelated but not independent
   (d) None of the above
5. What happens to the softmax function output as the temperature approaches 0 (assuming all inputs are different)?
   (a) It approaches a one-hot vector with 1 at the maximum input
   (b) It approaches a uniform distribution
   (c) All outputs approach 0.5
   (d) All outputs approach 1
6. What is the gradient of the quadratic form $z^T Az$ with respect to z, assuming that A is a symmetric matrix?
   (a) Az
   (b) $2Az$
   (c) $A^T z$
   (d) $z^T A$
7. What is the gradient of $\log\det A$ with respect to a symmetric positive definite matrix A?
   (a) A^{-1}
   (b) $-\frac{1}{2}A^{-1}$
   (c) A^T
   (d) $\log(A)A^{-1}$
8. What assumption does the Gaussian Mixture Model (GMM) make about the observations given the hidden state?
   (a) Each observation is sequentially dependent on the previous observation, following a Markovian process.

(b) Each observation follows a mixture of Poisson distributions rather than Gaussian distributions.

(c) Each observation is conditionally independent given the hidden state and follows a Gaussian distribution with state-dependent mean and covariance.

(d) The hidden states assign observations to clusters deterministically, without a probabilistic emission distribution.

9. What is the mathematical definition of the distortion function in K-means clustering?

(a) The sum of squared distances between each data point and its assigned cluster centroid.

(b) The sum of the absolute differences between each data point and its assigned cluster centroid.

(c) The determinant of the covariance matrix of the dataset.

(d) The Euclidean distance between all centroids in the dataset.

10. In a Hidden Markov Model (HMM) with M hidden states and D discrete observations, what are the key parameters that define the model?

(a) A transition probability matrix $T \in \mathbb{R}^{M \times M}$ and an emission probability matrix $E \in \mathbb{R}^{M \times D}$.

(b) An initial state probability vector $\pi \in \mathbb{R}^M$, a transition probability matrix $T \in \mathbb{R}^{M \times M}$ and an emission probability matrix $E \in \mathbb{R}^{M \times D}$.

(c) A transition probability matrix $T \in \mathbb{R}^{M \times M}$, an emission matrix $E \in \mathbb{R}^{M \times D}$, and a covariance matrix $\Sigma \in \mathbb{R}^{D \times D}$ capturing dependencies between observations.

(d) A single probability distribution over all possible hidden state sequences, rather than separate transition and emission matrices.

Latent Variable Models

Exercise 2. The Expectation-Maximization (EM) Algorithm

Consider a dataset consisting of n observations $x_1, \ldots, x_n$ where each observation is associated with a hidden variable $z_1, \ldots, z_n$. We assume that $(x_i, z_i)_{1 \leq i \leq n}$ are independent and identically distributed (i.i.d.) and follow a joint probability distribution parameterized by θ. The goal is to estimate θ by maximizing the marginal log-likelihood:

$$\log p_\theta(x) = \sum_{i=1}^{n} \log \sum_{z_i} p_\theta(x_i, z_i)$$

We introduce an auxiliary distribution $q(z_i)$ over the latent variables and use Jensen's inequality to derive a lower bound.

1. Show that the lower bound on the log-likelihood function is given by:

$$\log p_\theta(x) \geq \sum_{i=1}^{n} \sum_{z_i} q(z_i) \log \frac{p_\theta(x_i, z_i)}{q(z_i)}$$

2. Show that for all i, maximizing the lower bound with respect to $q(z_i)$ leads to:

$$q(z_i) = p_\theta(z_i | x_i)$$

where $p_\theta(z_i|x_i)$ is the posterior distribution of the latent variable given the observation.

3. Demonstrate that maximizing the lower bound with respect to θ is equivalent to maximizing the expectation of the complete log-likelihood function, where the expectation is taken with respect to the posterior distribution $p_\theta(z_i|x_i)$.

4. Write the Expectation-Maximization (EM) algorithm, detailing each step in a structured algorithmic format, including the E-step and the M-step.

Exercise 3. Gaussian Mixture Models (GMMs)

Consider a dataset consisting of n observations $x_1, \ldots, x_n$ in $\mathbb{R}^p$. We assume that the data is generated from a mixture of K Gaussian distributions, where each observation x_i has an associated latent variable z_i indicating which Gaussian component it belongs to. We make the following assumptions:

- For all observations $i \in \{1, \ldots, n\}$, the latent variable z_i follows a multinomial distribution:
$$z_i \sim \mathcal{M}(1, \pi_1, \ldots, \pi_K)$$
- For all clusters $j \in \{1, \ldots, K\}$, given that the latent variable z_i indicates cluster j, the observation x_i follows a Gaussian distribution:
$$x_i | z_i = j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

The model parameters are $\theta = (\pi, \mu, \Sigma)$. Our goal is to estimate these parameters using the Expectation-Maximization (EM) algorithm.

1. For all i and all j, compute the posterior distribution of the latent variable z_i, given the observation x_i and current parameter estimates:

$$p_\theta(z_i = j | x_i) = \frac{\pi_j \mathcal{N}(x_i | \mu_j, \Sigma_j)}{\sum_{j'=1}^{K} \pi_{j'} \mathcal{N}(x_i | \mu_{j'}, \Sigma_{j'})}$$

2. Compute $\mathbb{E}[\log p_\theta(X, Z)]$ with respect to the posterior distribution of the latent variables.

3. Prove that the update equations for the parameters π_j, μ_j, and Σ_j are given by:

$$\pi_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} p_{\theta^{(t)}}(z_i = j|x_i)$$
$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} p_{\theta^{(t)}}(z_i = j|x_i)x_i}{\sum_{i=1}^{n} p_{\theta^{(t)}}(z_i = j|x_i)}$$
$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^{n} p_{\theta^{(t)}}(z_i = j|x_i)(x_i - \mu_j^{(t+1)})(x_i - \mu_j^{(t+1)})^T}{\sum_{i=1}^{n} p_{\theta^{(t)}}(z_i = j|x_i)}$$

4. Write the K-means clustering algorithm in a structured format, detailing each step of the process.

5. Explain how the K-means algorithm can be used to initialize the EM algorithm, particularly in estimating the initial values of π_j, μ_j, and Σ_j, for all j.

Exercise 4. Hidden Markov Models (HMMs)

Consider a sequence of T observations $X = (X_1, \ldots, X_T)$ in $\mathbb{R}^{T\times d}$. We assume that the sequence is generated by a Hidden Markov Model (HMM) with M possible hidden states. The hidden state at time t is denoted by H_t.

We make the following assumptions:

- The initial hidden state follows a multinomial distribution:
$$H_1 \sim \mathcal{M}(1, \pi_1, \ldots, \pi_M)$$
- The probability of transitioning from state h at time t to state h' at time $t+1$ is given by:
$$p(H_{t+1} = h'|H_t = h) = Q_{hh'}$$
- For all $h \in \{1, \ldots, M\}$, given the hidden state, the observation follows a Gaussian distribution:
$$X_t|H_t = h \sim \mathcal{N}_d(\mu_h, \Sigma_h)$$

The model parameters are $\theta = (\pi, Q, \mu, \Sigma)$. Our goal is to estimate these parameters using the Expectation-Maximization (EM) algorithm.

We introduce the following probabilities:

- **Filtering probabilities:** For all $t \in \{1, \ldots, T\}$ and $h \in \{1, \ldots, M\}$, define:
$$\xi(t, h) := p(H_t = h|X_1, \ldots, X_t)$$
representing the probability of being in hidden state h at time t given all past observations up to time t.
- We define the vector $\xi_t \in \mathbb{R}^M$ as:
$$\xi_t = (\xi(t, 1), \ldots, \xi(t, M))^T$$

- **Forward (alpha) variables:** For all $t \in \{1, \ldots, T\}$ and $h \in \{1, \ldots, M\}$, define:
$$\alpha(t, h) := p(X_1, \ldots, X_t, H_t = h)$$
representing the joint probability of being in state h at time t and having observed $X_1, \ldots, X_t$.
- We define the vector $\alpha_t \in \mathbb{R}^M$ as:
$$\alpha_t = (\alpha(t, 1), \ldots, \alpha(t, M))^T$$

- For all $t \in \{1, \ldots, T\}$, the **emission tensor** $\Gamma(t) \in \mathbb{R}^{M\times M}$ is defined as:
$$\Gamma(t) = \text{diag}(p(X_t|H_t = 1), \ldots, p(X_t|H_t = M))$$

- **Smoothing probabilities:** For all $t \in \{1, \ldots, T\}$ and $h \in \{1, \ldots, M\}$, define:
$$\psi(t, h) := p(H_t = h|X_1, \ldots, X_T)$$
representing the probability of being in hidden state h at time t given all observations.
- We define the vector $\psi_t \in \mathbb{R}^M$ as:
$$\psi_t = (\psi(t, 1), \ldots, \psi(t, M))^T$$

- **Joint smoothing probabilities:** For all $t \in \{1, \ldots, T-1\}$, $h \in \{1, \ldots, M\}$, and $h' \in \{1, \ldots, M\}$, define:
$$\phi(t, h, h') := p(H_t = h, H_{t+1} = h'|X_1, \ldots, X_T)$$
.

- **Backward (beta) variables:** For all $t \in \{1, \ldots, T\}$ and $h \in \{1, \ldots, M\}$, define:
$$\beta(t, h) := p(X_{t+1}, \ldots, X_T|H_t = h)$$
representing the probability of the future observations given the current hidden state.
- We define the vector $\beta_t \in \mathbb{R}^M$ as:
$$\beta_t = (\beta(t, 1), \ldots, \beta(t, M))^T$$

1. Show that the forward variables α_t satisfy the recursive relation:

$$\alpha_1 = \Gamma(1)\pi, \quad \forall t \geq 2, \quad \alpha_t = \Gamma(t)Q^T\alpha_{t-1}$$

Then, show that the filtering probabilities ξ_t can be expressed in terms of α_t as:

$$\xi_t = \frac{\alpha_t}{\mathbb{1}^T\alpha_t}$$

2. Show that the marginal log-likelihood of the observations

$$\log p_\theta(X) = \log p_\theta(X_1, \ldots, X_T)$$

can be expressed as a function of the alpha variables.

3. Show that β_t satisfies the recursion:

$$\beta_T = \mathbb{1}, \quad \forall t \leq T-1, \quad \beta_t = Q\Gamma(t+1)\beta_{t+1}$$

Then, express the smoothing probabilities as:

$$\psi_t = \frac{\alpha_t \circ \beta_t}{\mathbb{1}^T\alpha_T}, \quad \phi_t = \frac{\mathrm{diag}(\alpha_t)Q\Gamma(t+1)\mathrm{diag}(\beta_{t+1})}{\mathbb{1}^T\alpha_T}$$

4. Show that the expected value of the complete log-likelihood function with respect to the posterior distribution of the hidden states is given by:

$$\begin{aligned}\mathbb{E}_{H|X}[\log p_\theta(H, X)] = &\sum_{h=1}^{M} \log(\pi_h)\psi(1,h) \\ &+ \sum_{t=1}^{T-1}\sum_{h=1}^{M}\sum_{h'=1}^{M} \log Q_{hh'}\phi(t,h,h') \\ &+ \sum_{t=1}^{T}\sum_{h=1}^{M} \log \mathcal{N}(X_t|\mu_h, \Sigma_h)\psi(t,h).\end{aligned}$$

5. Show that after training the HMM, the next hidden state H_{T+1} given the sequence of observations $X = (X_1, \ldots, X_T)$ satisfies, for all $h \in \{1, \ldots, M\}$:

$$p(H_{T+1} = h|X_1, \ldots, X_T) = \sum_{h'=1}^{M} Q_{h'h}\xi(T, h').$$

Probability and Statistics

Exercise 5. We recall the definition of entropy for a discrete random variable X with probability mass function p_X taking values in $\mathcal{X}$:

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x)\log p_X(x)$$

For two discrete random variables X, Y with joint distribution $p_{X,Y}$ and marginals p_X, p_Y, the conditional entropy is defined as:

$$H(X \mid Y) = -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_Y(y)}$$

1. Show that $H(X \mid Y)$ can be written as:

$$H(X \mid Y) = \sum_{y\in\mathcal{Y}} p_Y(y)H(X \mid Y = y)$$

where $H(X \mid Y = y) = -\sum_{x\in\mathcal{X}} p_{X|Y=y}(x \mid y)\log p_{X|Y=y}(x \mid y)$ is the entropy of the conditional distribution of X given $Y = y$.

2. Prove that:

$$H(X,Y) = H(Y) + H(X \mid Y)$$

Hint: Start by writing out the definition of $H(X,Y)$ and use the chain rule for probabilities.

3. The mutual information between X and Y is defined as $I(X;Y) = H(X) - H(X \mid Y)$. Show that it is symmetric, i.e.:

$$H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

Hint: Use the result from part 2.

Exercise 6. Let X be a Poisson random variable with parameter $\lambda > 0$, taking values in $\mathbb{N}$ (the set of non-negative integers). Its probability mass function is:

$$p_\lambda(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x \in \mathbb{N}$$

We recall the power series expansion of the exponential:

$$e^t = \sum_{k=0}^{\infty}\frac{t^k}{k!}$$

1. For N iid observations $X_1, ..., X_N$ from a Poisson(λ), write the likelihood function $\mathcal{L}(\lambda)$.

2. Take the log and compute $\frac{d}{d\lambda} \log \mathcal{L}(\lambda)$. Deduce the Maximum Likelihood Estimator (MLE) $\hat{\lambda}_N$.

3. Using the power series expansion above:

   (a) Show that $\sum_{x=0}^{\infty} p_\lambda(x) = 1$

   (b) Show that $\mathbb{E}[X] = \lambda$

4. Using similar techniques, calculate $\mathbb{E}[X^2]$ and deduce $\mathrm{Var}(X)$.

Exercise 7. Consider a simple neural network that takes as input a vector $x \in \mathbb{R}^2$ and applies the following transformations: 1. Multiplication by a matrix $A \in \mathbb{R}^{3\times 2}$ 2. Element-wise sigmoid activation: $\sigma(z) = \frac{1}{1+e^{-z}}$

For a single input x and target $y \in [0,1]^3$, let:

$$z = Ax \in \mathbb{R}^3$$
$$\hat{y} = \sigma(z) \in \mathbb{R}^3 \quad \text{(element-wise sigmoid)}$$
$$L(A) = \frac{1}{2}\|\hat{y} - y\|_2^2 = \frac{1}{2}\sum_{i=1}^{3}(\hat{y}_i - y_i)^2$$

Compute $\nabla_A L$, the gradient of L with respect to A (a 3×2 matrix). Follow these steps:

1. Verify that $\frac{d}{dz}\sigma(z) = \sigma(z)(1 - \sigma(z))$

2. Write the jacobian $\frac{\partial L}{\partial \hat{y}}$ (a row vector in $\mathbb{R}^3$)

3. Using the chain rule, express the jacobian $\frac{\partial L}{\partial z}$ in terms of $\frac{\partial L}{\partial \hat{y}}$ and $\frac{\partial \hat{y}}{\partial z}$

4. Finally, express $\nabla_A L$ in terms of $\frac{\partial L}{\partial z}$ and x