Final Project Theory Track Questions
Applied AI for Public Health

Fall 2025

Instructions

To successfully complete the theory track, attempt as many questions as you can and submit good-faith solutions to your strongest five. **You only need to turn in five solved questions (good-faith attempts) out of the ten to pass.** Show your reasoning clearly; partial progress earns partial credit. Unless otherwise stated, you may assume all random variables are well-behaved so that expectations and variances exist.

1. ℓ_2 **distance and cosine similarity.** Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ with $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$. Prove that minimizing the squared Euclidean distance $\|\mathbf{x}-\mathbf{y}\|_2^2$ is equivalent to maximizing the cosine similarity $\mathbf{x}^\top \mathbf{y}$. (Hint: expand the squared norm and use the unit-norm assumption.)

2. **Optimal regression function.** Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. Show that the function f which minimizes the expected squared loss $\mathbb{E}\big[(Y - f(X))^2\big]$ is the conditional expectation $f^*(x) = \mathbb{E}[Y \mid X = x]$. (Hint: condition on X, expand the square, and use the tower property to compare any f to f^*.)

3. **Convexity and global minima.** Recall that a function $g : \mathbb{R} \to \mathbb{R}$ is convex if $g(\lambda x + (1-\lambda)y) \le \lambda g(x) + (1-\lambda)g(y)$ for all x, y and $\lambda \in [0, 1]$. Prove that any local minimum of a differentiable convex function (you may treat the one-dimensional case for simplicity) is also a global minimum. Explain where you use first-order optimality conditions (i.e., $\nabla g(x_{\text{loc}}) = 0$) in your argument.

4. **Bias–variance decomposition.** Let $\hat{f}(X)$ be an estimator for the target Y and suppose $Y = f(X) + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] = 0$ and $\mathrm{Var}(\varepsilon \mid X) = \sigma^2$. Show that

$$\mathbb{E}\big[(Y - \hat{f}(X))^2\big] = \mathrm{Bias}[\hat{f}(X)]^2 + \mathrm{Var}[\hat{f}(X)] + \sigma^2,$$

   where the bias and variance are taken with respect to the randomness in the training data used to fit $\hat{f}$. Clearly label each conditioning step and define the bias term you use.

5. **InfoNCE as cross-entropy.** Recall the InfoNCE objective for contrastive learning with one positive pair $(\mathbf{v}, \mathbf{v}^+)$ and K negative examples $\{\mathbf{v}_k^-\}_{k=1}^K$:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(s(\mathbf{v}, \mathbf{v}^+)/\tau)}{\exp(s(\mathbf{v}, \mathbf{v}^+)/\tau) + \sum_{k=1}^K \exp(s(\mathbf{v}, \mathbf{v}_k^-)/\tau)},$$

where $s(\cdot, \cdot)$ is a similarity score and τ is a temperature parameter. Show that optimizing this loss is equivalent to minimizing the cross-entropy loss of a $(K+1)$-class classifier that predicts which candidate is the true positive (one class per candidate). Explicitly construct the classifier viewpoint by (i) defining the logits for each class and (ii) mapping InfoNCE terms to the softmax probabilities and label indicator.

6. **TF–IDF attenuates stopwords.** Suppose documents $\mathbf{x}$ and $\mathbf{y}$ are represented with TF–IDF weights $x_w = \mathrm{tf}_x(w) \cdot \mathrm{idf}(w)$ and $y_w = \mathrm{tf}_y(w) \cdot \mathrm{idf}(w)$. Here $\mathrm{tf}_x(w)$ is the term frequency of word w in document x, $\mathrm{df}(w)$ is the document frequency, N is the number of documents, and $\mathrm{idf}(w) = \log \frac{N}{\mathrm{df}(w)}$. Show that the dot product $\mathbf{x}^\top \mathbf{y} = \sum_w \mathrm{tf}_x(w)\ \mathrm{tf}_y(w)\ \mathrm{idf}(w)^2$ weights each word by $\mathrm{idf}(w)^2$. Use this algebra to interpret why high document-frequency words (stopwords) contribute little to similarity.

7. **Gradient descent on a 1D quadratic.** Recall that gradient descent updates parameters by $w_{t+1} = w_t - \eta \nabla f(w_t)$. Consider $f(w) = \frac{1}{2} a (w - w^*)^2$ with $a > 0$.

   (a) Derive the gradient descent update $w_{t+1} = w_t - \eta \nabla f(w_t)$ and express the error $e_t = w_t - w^*$ as a linear recurrence.

   (b) Show that $|e_t|$ converges to zero if and only if $0 < \eta < 2/a$. (Hint: analyze the magnitude of the recurrence ratio.)

8. **KL divergence between Gaussians.** Recall that the Kullback–Leibler divergence from $\mathcal{N}(\mu_0, \sigma_0^2)$ to $\mathcal{N}(\mu_1, \sigma_1^2)$ is defined as

$$D_{\mathrm{KL}}\big(\mathcal{N}(\mu_0, \sigma_0^2) \parallel \mathcal{N}(\mu_1, \sigma_1^2)\big) = \int \log \frac{p_0(x)}{p_1(x)}\, p_0(x)\, dx.$$

   Derive the standard closed-form expression for this divergence in one dimension and explain why it is always non-negative. (Optional: note which metric properties D_{KL} fails.)

9. **Softmax temperature intuition.** Let $\mathbf{z} \in \mathbb{R}^m$ be logits for a categorical distribution and define the tempered softmax $p_i(\tau) = \dfrac{\exp(z_i/\tau)}{\sum_{j=1}^m \exp(z_j/\tau)}$.

   (a) Evaluate the limits of $p_i(\tau)$ as $\tau \to 0^+$ and as $\tau \to \infty$ for fixed logits.

   (b) Illustrate the effect of temperature by analyzing a three-class example with logits $(2, 1, 0)$ and plotting or tabulating $p_i(\tau)$ for $\tau \in \{0.1, 0.5, 1, 2, 5\}$. Explain how decreasing versus increasing τ sharpens or flattens the distribution.

10. **Choose-your-own fact.** Identify one mathematical or statistical result related to the course material that you find interesting (e.g., a theorem, lemma, or algorithmic guarantee). (a) State the result precisely. (b) Explain why it matters for machine learning. (c) Outline a proof sketch or intuitive justification that would convince a peer.