

CS280 Fall 2018 Assignment 1

Part A

ML Background

Due in class, October 12, 2018

Name:

Student ID:

1. MLE (5 points)

Given a dataset $\mathcal{D} = \{x_1, \dots, x_n\}$. Let $p_{emp}(x)$ be the empirical distribution, i.e., $p_{emp}(x) = \frac{1}{n} \sum_{i=1}^n \delta(x, x_i)$ and let $q(x|\theta)$ be some model.

- Show that $\arg \min_q KL(p_{emp}||q)$ is obtained by $q(x) = q(x; \hat{\theta})$, where $\hat{\theta}$ is the Maximum Likelihood Estimator and $KL(p||q) = \int p(x)(\log p(x) - \log q(x))dx$ is the KL divergence.

2. Properties of l_2 regularized logistic regression (10 points)

Consider minimizing

$$J(\mathbf{w}) = -\frac{1}{|D|} \sum_{i \in D} \log \sigma(y_i \mathbf{x}_i^T \mathbf{w}) + \lambda \|\mathbf{w}\|_2^2$$

where $\lambda > 0$ and $y_i \in -1, +1$. Answer the following true/false questions and **explain why**.

- $J(\mathbf{w})$ has multiple locally optimal solutions: T/F?
- Let $\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} J(\mathbf{w})$ be a global optimum. $\hat{\mathbf{w}}$ is sparse (has many zeros entries): T/F?

3. Gradient descent for fitting GMM (15 points)

Consider the Gaussian mixture model

$$p(\mathbf{x}|\theta) = \sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

Define the log likelihood as

$$l(\theta) = \sum_{n=1}^N \log p(\mathbf{x}_n|\theta)$$

Denote the posterior responsibility that cluster k has for datapoint n as follows:

$$r_{nk} := p(z_n = k|\mathbf{x}_n, \theta) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \Sigma_{k'})}$$

- Show that the gradient of the log-likelihood wrt μ_k is

$$\frac{d}{d\mu_k} l(\theta) = \sum_n r_{nk} \Sigma_k^{-1} (\mathbf{x}_n - \mu_k)$$

- Derive the gradient of the log-likelihood wrt π_k without considering any constraint on π_k . (bonus: with constraint $\sum_k \pi_k = 1$.)
- Derive the gradient of the log-likelihood wrt Σ_k without considering any constraint on Σ_k . (bonus: with constraint Σ_k be a symmetric positive definite matrix.)