Midterm (Group A)

Introduction to Machine Learning
Fall 2024
Instructor: Anna Choromanska

Problem 1 (10 points)

Suppose we have m boxes. i^{th} box ($i = \{1, 2, \ldots, m\}$) contains n_i apples and r_i oranges. One of the boxes is chosen at random (with equal probability of choosing any box) and an item is selected from the box and found to be an apple. Use Bayes' rule to find the probability that the apple came from the j^{th} box ($j \in \{1, 2, \ldots, m\}$).

Problem 2 (10 points)

Show the first two iterations (after the initialization) of the k-means clustering algorithm (show centers and assignments of data points to clusters) for the following 2D data set: $(-3,-1),(-1,-3),(-2,-6),(-5,-7),(3,1)$, $(2,3),(3,6),(8,1)$. Assume the number of centers is equal to 2 and the centers are initialized to $(-4,-5)$ and $(5,4)$.

Problem 3 (20 points)

Suppose you have a 2-class classification problem, where each class is Gaussian. Let $\theta = \{\alpha, \mu_1, \Sigma_1, \mu_2, \Sigma_2\}$ denote the set of model parameters. Suppose the class probability $p(y|\theta)$ is modelled via the Bernoulli distribution, i.e. $p(y|\theta) = \alpha^y(1-\alpha)^{1-y}$, and the probability of the data $p(x|y,\theta)$ is modelled as $p(x|y,\theta) = \mathcal{N}(x|\mu_y, \Sigma_y)$. Recover the parameters of the model from maximum likelihood approach [2 point for α, 3 points for μ's, and 5 points for Σ's]. Assume the data are i.i.d. and N is the number of data samples. Show all derivations. Next, suppose you want to make a classification decision by assigning proper label y to a given data point x. You decide the label based on Bayes optimal decision $y = \arg\max_{\hat{y}=\{0,1\}} p(\hat{y}|x)$. Prove that the decision boundary is quadratic when covariances Σ_1 and Σ_2 are not equal [10 points].

Problem 4 (10 points)

A kernel is an efficient way to write out an inner product between two feature vectors computed from a pair of input vectors as follows:

$$K(x, z) = \phi(x)^\top \phi(z).$$

Assume that both inputs are 2-dimensional. Show the explicit features mapping ϕ for a kernel given as follows:

$$K(x, z) = (x^\top z + c)^2.$$

Problem 5 (15 points)

What is the VC dimension of the hypothesis space consisting of triangles in the 2D plane (justify your answer)? Points inside the triangle are classified as positive examples.

Problem 6 (20 points)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. samples from a population with pdf

$$f(x|\theta) = \frac{1}{\theta} x^{(1-\theta)/\theta}, \quad 0 < x < 1, \quad 0 < \theta < \infty.$$

Find the maximum likelihood estimator for θ (15 points). Calculate an estimate using this estimator when $x_1 = 0.10,\ x_2 = 0.22,\ x_3 = 0.54,\ x_4 = 0.36$ (5 points).

Problem 7 (15 points)

Consider a linear regression problem in which we want to weight N different training examples differently. Specifically, we want to minimize:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{N} w_i (\theta^\top x_i - y_i)^2,$$

where the w_i and y_i are just scalars and w_i are non-negative. Also, assume vectors θ and x_i are D-dimensional.

a) (10 points) Show that $J(\theta)$ can also be written as $J(\theta) = (X\theta - y)^\top W (X\theta - y)$. Specify the sizes and specify the entries of the matrices W and X and the vector y and do so using scalars w_i, y_i, and entries of $x_i = [x_i(1) \;\; x_i(2) \;\; \ldots x_i(D)]^\top$.

b) (5 points) Suppose we have a training set of N independently distributed examples: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ and wish to learn the conditional distribution:

$$p(y_i | x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-(y_i - \theta^\top x_i)^2 / (2\sigma_i^2)\right)$$

by maximizing the conditional log-likelihood: $\sum_{i=1}^{N} \log p(y_i | x_i; \theta)$. Show that finding the maximum conditional likelihood estimate of θ reduces to solving a weighted linear regression problem. State clearly what the w_i values should be in terms of the σ_i values.