Midterm (Group A)

Introduction to Machine Learning
Fall 2018
Instructor: Anna Choromanska

Problem 1 (45 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 [5 points for α, 10 points for μ's, and 15 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 linear when covariances Σ_1 and Σ_2 are equal and otherwise the boundary is quadratic. [5 points for each case] Illustrate on 2d example the decision boundary for the case when covariances are not equal clearly indicating which class is more concentrated around its mean [5 points].

Problem 2 (35 points)

Consider 2d family of classifiers given by axis-aligned rectangles. What is the VC dimension of this family?

Problem 3 (35 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) (20 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) (15 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.

Problem 4 (35 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:

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

Assume that both inputs are 2-dimensional and write out the explicit mapping ϕ that mimics the kernel value for a 3$^{\text{rd}}$-order polynomial kernel as follows:

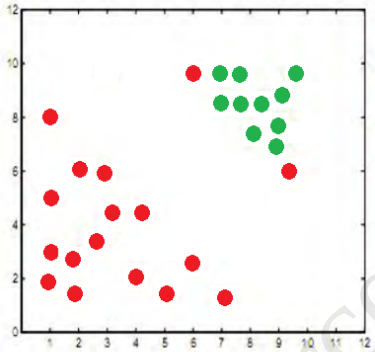
$$K(x, y) = (x^\top y + 1)^3.$$

Problem 5 (35 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: $(4, 2), (0, -1), (1, 4), (2, 8), (3, 5), (8, 8), (3, 3),$ $(10, 10), (20, 18)$, and $(12, 9)$. Assume the number of centers is equal to 2 and the centers are initialized to $(1, 1)$ and $(7, 8)$.

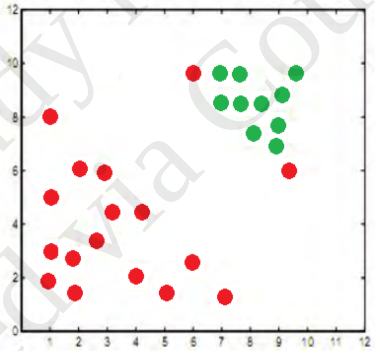
Problem 6 (30 points)

The original SVM proposed was a linear classifier. In order to make SVM non-linear we map the training data on to a higher dimensional feature space and then use a linear classifier in the that space. This mapping can be done with the help of kernel functions. For this question assume that we are training an SVM with a quadratic kernel - i.e. our kernel function is a polynomial kernel of degree 2. This means the resulting decision boundary in the original feature space may be parabolic in nature. The dataset on which we are training is given below: The

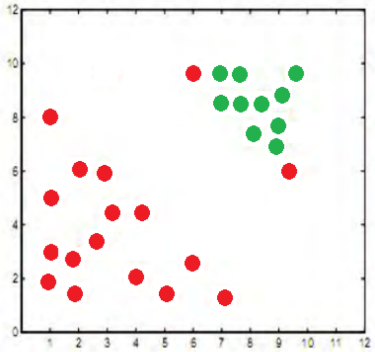


slack penalty C will determine the location of the separating parabola. Please answer the following questions qualitatively.

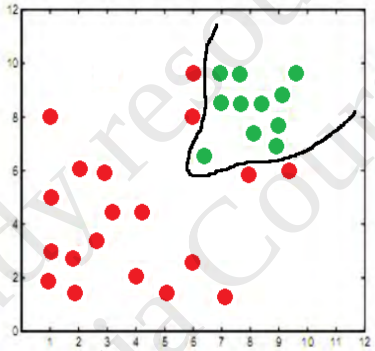
a) [10 points] Where would the decision boundary be for very large values of C? (Remember that we are using a quadratic kernel). Justify your answer in one sentence and then draw the decision boundary in the figure below.



b) [10 points] Where would the decision boundary be for C nearly equal to 0? Justify your answer in one sentence and then draw the decision boundary in the figure below.



c) [10 points] Now suppose we add three more data points as shown in figure below. Now the data are not quadratically separable, therefore we decide to use a degree-5 kernel and find the following decision boundary. Most probably, our SVM suffers from a phenomenon which will cause wrong classification of new data points. Name that phenomenon, and in one sentence, explain what it is.



Problem 7 (35 points)

Consider the discrete distribution $\{p_k | k = 1, 2, \ldots, N\}$. The entropy of this distribution is given as $H = -\sum_{k=1}^{N} p_k \log p_k$. What is the distribution that maximizes this entropy? Show formal derivations using the method of Lagrange multipliers.