CS534 — Written Homework Assignment 2 — Due Oct 20th 11:59pm, 2018

Please submit electronically via TEACH. Your submission should be a single PDF file.

1. (Maximum likelihood estimation.) In DNA, also known as the Code of Life, there exist four different possible bases: adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). We are given an organism of unknown DNA base frequencies. Let p_a, p_c, p_g, and p_t be those unknown frequencies. Assume that we have obtained a strand of DNS sequences and we want to estimate the unknown frequencies. Let n_a, n_c, n_g, n_t be the corresponding number of bases that you observe for A, C, T and G respectively. Please derive the maximum likelihood estimates for the unknown parameters p_a, p_c, p_g, and p_t.
   Hint: it it important to remember that $p_a + p_c + p_g + p_t = 1$. You can incorporate this constraint into the optimization using the lagrangian. If you are not familiar with this concept, here is a blog post that gives a good explanation
   `https://medium.com/@andrew.chamberlain/a-simple-explanation-of-why-lagrange-multipliers-works-253e2cdcbf74`.

2. (Naive Bayes Classifier) Consider the following training set:

A	B	C	Y
0	1	1	0
1	1	1	0
0	0	0	0
1	1	0	1
0	1	0	1
1	0	1	1

   (a) Learn a Naive Bayes classifier by estimating all necessary probabilities.
   (b) Compute the probability $P(y = 1|A = 1, B = 0, C = 0)$.
   (c) Suppose we know that A, B and C are independent random variables, can we say that the Naive Bayes assumption is valid? (Note that the particular data set is irrelevant for this question). If your answer is yes, please explain why; if you answer is no please give an counter example.

3. (Maximum A Posterior Estimation.) As discussed in class, consider using a beta prior $Beta(2, 2)$ for estimating p, the probability of head for a weighted coin. What is the posterior distribution of p after we observe 5 coin tosses and 2 of them are head? What is the posterior distribution of p after we observe 50 coin tosses and 20 of them are head? Plot the pdf function of these two posterior distributions. Assume that $p = 0.4$ is the true probability, as we observe more and more coin tosses from this coin, what do you expect to happen to the posterior?

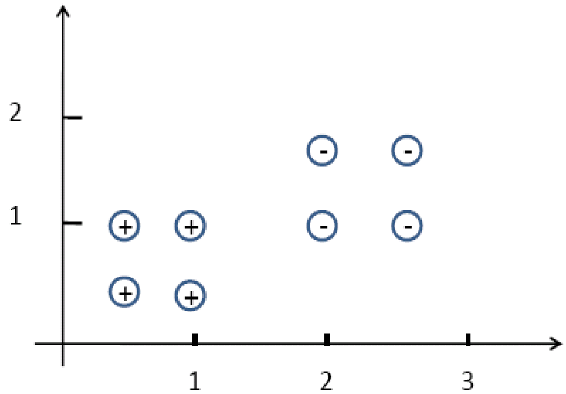
4. (Perceptron) The perceptron algorithm will only converge if the data is linearly separable. It is possible to *force* your data to be linearly separable as follows. If you have N data points in D dimensions, map data point $\vec{x}_n$ to the $(D + N)$-dimensional point $\langle \vec{x}_n, e_n \rangle$, where e_n is a N-dimensional vector of all zeros but one 1 at the nth position. (Eg., $e_4 = \langle 0, 0, 0, 1, 0, \ldots \rangle$.)

   (a) Show that if you apply this mapping the data becomes linearly separable (you may wish to do so by providing a weight vector $\vec{w}$ in $(D + N)$-dimensional space that successfully separates the data).
   (b) How does this mapping affect generalization?

5. (Kernels) Cubic Kernels. In class, we showed that the quadratic kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^2$ was equivalent to mapping each $\mathbf{x} = (x_1, x_2) \in R^2$ into a higher dimensional space where

$$\Phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1).$$

   Now consider the cubic kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^3$. What is the corresponding Φ function?



6. (Linear SVM) Apply linear SVM without soft margin to the following problem. Note that the two right most positive points are $(1, 0.5)$ and $(1, 1)$. The two left most negative points are $(2, 1)$ and $(2, 1.5)$.

   a. Please mark out the support vectors, the decision boundary ($\mathbf{w}^T\mathbf{x} + b = 0$) and $\mathbf{w}^T\mathbf{x} + b = 1$ and $\mathbf{w}^T\mathbf{x} + b = -1$. Note that you don't need to solve the optimization problem for this, just eyeball the solution.

   b. Please solve for $\mathbf{w}$ and b based on the support vectors you identified in (a).