

CS 281A/Stat 241A Homework Assignment 4 (due October 28)

1. A useful Kullback-Leibler decomposition.

Given distributions $p(x_A, x_B)$ and $q(x_A, x_B)$, show the following:

$$D(p(x_A, x_B) \parallel q(x_A, x_B)) = D(p(x_A) \parallel q(x_A)) + \sum_{x_A} p(x_A) D(p(x_B | x_A) \parallel q(x_B | x_A)).$$

2. An alternative form of iterative proportional fitting.

Derive Eq. (9.72) in the text.

3. Classification.

The course website contains a data set “classification.dat” of (x_n, y_n) pairs, where the x_n are 2-dimensional vectors and y_n is a binary label.

- Plot the data, using 0's and X's for the two classes. The plots in the following parts should be plotted on top of this plot.
- Fit a generative model to the data, using Gaussian class-conditional densities with equal covariance matrices. Calculate the posterior probability of class 1, and plot the line where this probability is equal to 0.5.
- Write a program to fit a logistic regression model using the IRLS algorithm (remembering to include the intercept term). Plot the line where the logistic function is equal to 0.5.
- Write a program to fit a logistic regression model using stochastic gradient ascent. Plot the line where the logistic function is equal to 0.5.
- Fit a linear regression to the problem, treating the class labels as real values 0 and 1. (You can solve the linear regression in any way you'd like, including solving the normal equations, using the LMS algorithm, or calling the built-in routines in Matlab or Splus). Plot the line where the linear regression function is equal to 0.5.
- The data set “classification.test” is a separate data set generated from the same source. Test your fits from parts (b), (c), (d) and (e) on these data and compare the results.
- (Gedanken). Suppose that you randomly delete half of the points in class 0 for “classification.dat” and refit the models. What do you think will happen to the line in part (b)? In part (c)?

4. Invariance of ML estimates.

Let $p(x | \theta)$ be a probability model, where the parameter θ lies in some set Ω . Let $\theta = f(\eta)$ be one-to-one and onto, so that η is also a parameter. Consider an IID data set. Prove

that maximum likelihood estimates are invariant to one-to-one reparameterization. That is, prove that $\hat{\theta}_{ML} = f(\hat{\eta}_{ML})$.

5. Concavity of log likelihood in the mixing proportions.

Given k fixed distributions $p_1(x), p_2(x), \dots, p_k(x)$, consider the problem of fitting the mixing proportions $\theta = (\pi_1, \pi_2, \dots, \pi_k)$, where $\sum_i \pi_i = 1$, $\pi_i \geq 0$, in the mixture distribution:

$$p(x|\theta) = \pi_1 p_1(x) + \pi_2 p_2(x) + \dots + \pi_k p_k(x).$$

Prove that, given N IID samples of X , the log likelihood $\ell(\theta|\mathcal{D})$ is concave in θ .

6. K-means derivation.

Derive the K-means algorithm:

$$z_n^i = \begin{cases} 1 & \text{if } i = \arg \min_j \|x_n - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$
$$\mu_i = \frac{\sum_n z_n^i x_n}{\sum_n z_n^i}.$$

as an algorithm that performs coordinate descent in the following cost (“distortion”) function:

$$J = \sum_{n=1}^N \sum_{i=1}^K z_n^i \|x_n - \mu_i\|^2.$$