DS 215: Assignment 3

Full Marks: 100

Problem 1: Consider a K-component Gaussian Mixture Model (GMM) where the covariance for each Gaussian is identical and known as σ^2I. Note that σ^2 is already observed and not a parameter to be estimated. Show that for $\sigma \to 0$, the EM algorithm to estimate GMM parameters coincides with K-means clustering.

15

Problem 2: In a course, a student can get one of the four possible grades: A, B, C and D. The probability for each student getting either of these grades are: $P(A) = \frac{1}{2}$, $P(B) = \mu$, $P(C) = 2\mu$ and $P(D) = \left(\frac{1}{2} - 3\mu\right)$. We then observe some data: let the number of students in the class who got A, B, C or D are a, b, c and d respectively.

(a) What is the ML estimate for μ given all these observations?

(b) Now consider a different scenario. In the data we observe, we no longer observe a or b separately. Rather we observe h, which is the total number of students who got either A or B. So, in this case, we do not know a or b, but we observe and know h (which $= a + b$). We still observe c and d separately as before. We now intend to use EM algorithm to find an estimate of μ. Therefore, given all these information in part **b**, find:

E-step: Given $\hat{\mu}$, the current estimate of μ, what are the expected values of a and b?
M-Step: Given $\hat{a}$ and $\hat{b}$, the expected values of a and b, what is the ML estimate for μ?

10+15 = 25

Problem 3: Suppose you are given 30 data samples in the form of $\{(x_i, y_i)\}_{i=1\cdots30}$ and you are trying to find the best fit line for these points via linear regression.

(a) The original data samples $\{(x_i, y_i)\}_{i=1\cdots30}$ are not revealed to you, but instead you are provided with some statistics of data:

$$\begin{aligned} \bar{x} = 4, \quad \bar{y} = 3 & \qquad \text{(sample means)} \\ \sigma_x^2 = 2.25, \quad \sigma_y^2 = 0.25 & \qquad \text{(sample variances)} \\ \rho_{xy} = 0.7 & \qquad \text{(sample correlation coefficient)} \end{aligned}$$

Derive the equation of the best-fit line.

(b) Does the best-fit line pass through the point $(\bar{x}, \bar{y})$?

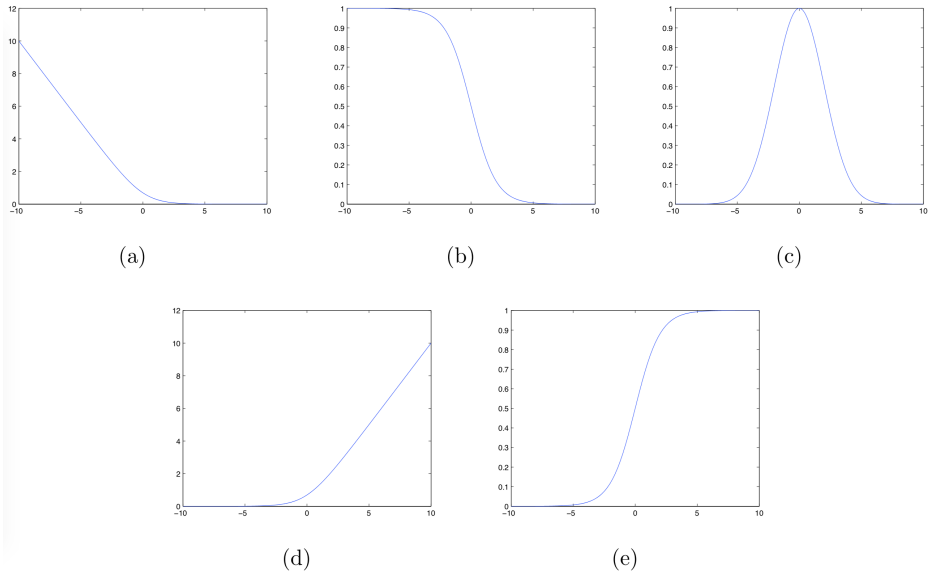
16+4 = 20

Problem 4: Assume $y_i \sim N\left(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2\right), i = 1, 2, \cdots, N$ and $\boldsymbol{\beta} \in \mathbb{R}^p$. The parameters $\beta_j, j = 1, \cdots, p$ are each distributed as $N\left(0, \tau^2\right)$, independently of one another. Assuming σ^2 and τ^2 are known, and β_0 is not governed by a prior (or has a flat improper prior), show that the (minus) log-posterior density of $\boldsymbol{\beta}$ is proportional to $\sum_{i=1}^N \left(y_i - \beta_0 - \sum_j x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^p \beta_j^2$, where $\lambda = \sigma^2/\tau^2$.

20

Problem 5: Generally speaking, a classifier can be written as $H(x) = \text{sign}(F(x))$, where $H(x) : \mathbb{R}^d \to \{-1, 1\}$ and $F(x) : \mathbb{R}^d \to \mathbb{R}$. To obtain the parameters in $F(x)$, we need to minimize the loss function averaged over the training set: $\Sigma_i L(y^i F(x^i))$. Here L is a function of $yF(x)$. For example, for linear classifiers, $F(x) = w_0 + \Sigma_{j=1}^d w_j x_j$ and $yF(x) = y(w_0 + \Sigma_{j=1}^d w_j x_j)$.

(a) Which loss functions below are appropriate to use in classification? For the ones that are not appropriate, explain why not. In general, what conditions do L have to satisfy in order to be an appropriate loss function? The x axis is $yF(x)$ and the y axis is $L(yF(x))$.



(a) (b) (c)

(d) (e)

(b) Of the above loss functions appropriate to use in classification, which one is the most robust to outliers? Justify your answer.

(c) Let $F(x) = w_0 + \Sigma_{j=1}^{d} w_j x_j$ and $L(yF(x)) = \frac{1}{1+\exp(yF(x))}$. Suppose you use gradient descent to obtain the optimal parameters w_0 and w_j. Give the update rules for these parameters.

8+5+7 = 20