CSE 569 Homework #5

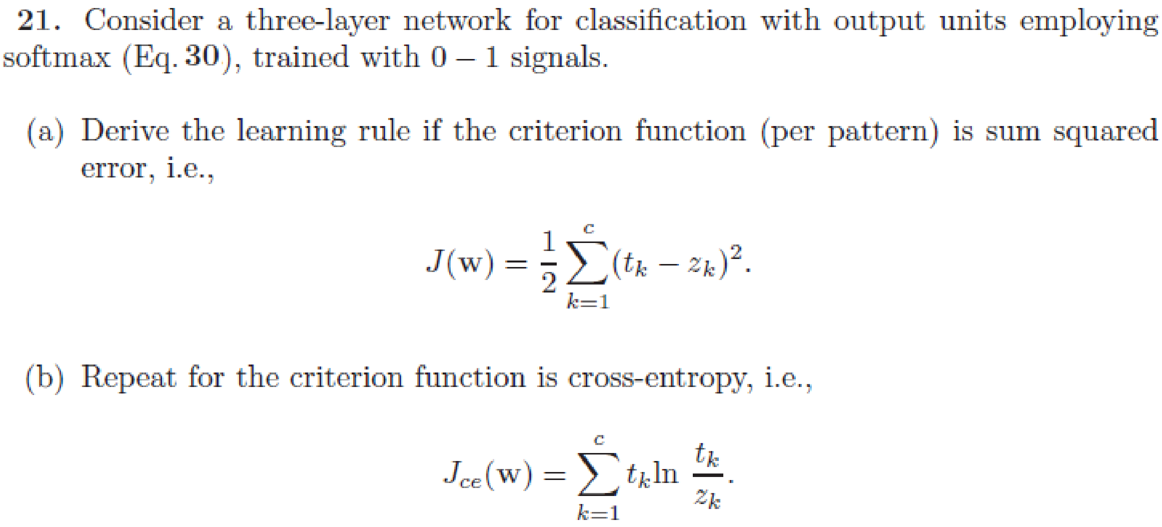
Total 3 points

Due Tuesday, December 1 **by 11:59pm**.

Question 1. (Based on Problem 6 of Chapter 6)
One might argue that the backpropagation learning rule should make the weight update *inversely* relate to *f'*(net)—that is, the weight update should be large where the output does not vary much. In fact, as shown in Eqn. (17) of the textbook (the first formula on Slide 16 of the lecture notes), the learning rule is linear in *f'*(net). Explain intuitively why the learning rule should be linear in *f'*(net).

Question 2. From Slides 32-34 of 08-Multilayer-Neural-Networks, we saw that the outputs of the nodes on the last layer approximate the posteriori probabilities. The derivation in the slides has omitted a few steps (especially for $\lim_{n\to\infty}\frac{1}{n}J_k(\boldsymbol{w})$). Fill in those steps.

Question 3. **(Problem 21 of Chapter 6)**



21. Consider a three-layer network for classification with output units employing softmax (Eq. **30**), trained with $0-1$ signals.

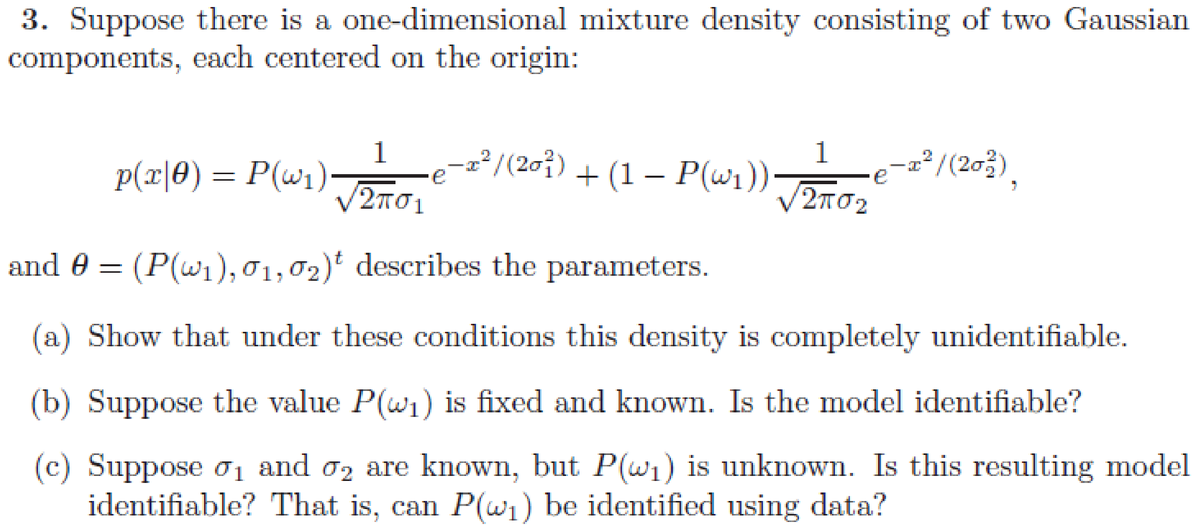
(a) Derive the learning rule if the criterion function (per pattern) is sum squared error, i.e.,

$$J(\mathrm{w}) = \frac{1}{2}\sum_{k=1}^{c}(t_k - z_k)^2.$$

(b) Repeat for the criterion function is cross-entropy, i.e.,

$$J_{ce}(\mathrm{w}) = \sum_{k=1}^{c} t_k \ln\frac{t_k}{z_k}.$$

Question 4. (Problem 3 of Chapter 10)



3. Suppose there is a one-dimensional mixture density consisting of two Gaussian components, each centered on the origin:

$$p(x|\boldsymbol{\theta}) = P(\omega_1)\frac{1}{\sqrt{2\pi}\sigma_1}e^{-x^2/(2\sigma_1^2)} + (1 - P(\omega_1))\frac{1}{\sqrt{2\pi}\sigma_2}e^{-x^2/(2\sigma_2^2)},$$

and $\boldsymbol{\theta} = (P(\omega_1), \sigma_1, \sigma_2)^t$ describes the parameters.

(a) Show that under these conditions this density is completely unidentifiable.

(b) Suppose the value $P(\omega_1)$ is fixed and known. Is the model identifiable?

(c) Suppose σ_1 and σ_2 are known, but $P(\omega_1)$ is unknown. Is this resulting model identifiable? That is, can $P(\omega_1)$ be identified using data?

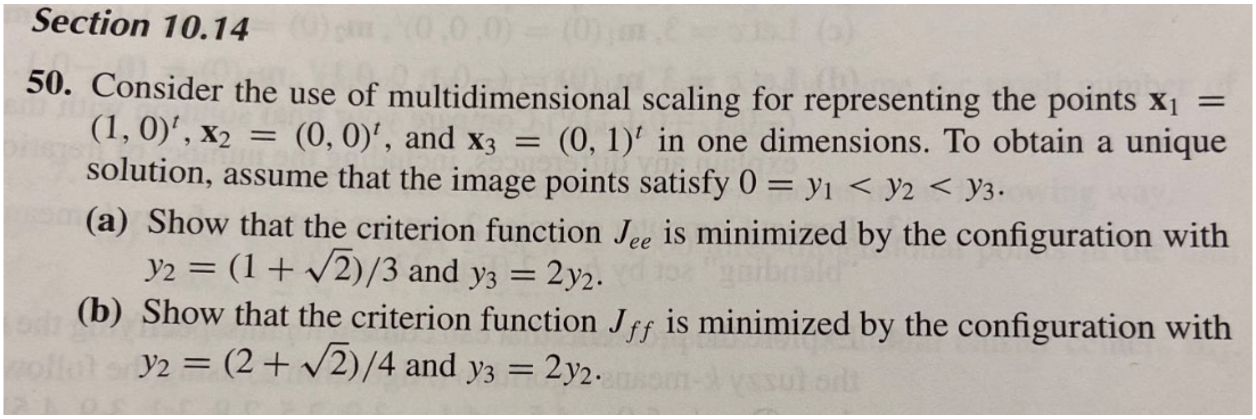
Question 5. Consider the *k*-Means algorithm from Slide 16 of the notes (or Algorithm 1 in Section 10.4.3 of the textbook):

(1) Will the algorithm still work, if you initialize all $\boldsymbol{\mu}_i$, *i*=1,…,*C*, to the same initial value?

(2) What is the complexity of the algorithm, in terms of *N* (the number of samples), *C* (the number of clusters), and *T* (the total number of iterations until convergence)?

(3) Consider the case of *N* data points drawn from a mixture models with *C* normal densities with means at $\boldsymbol{\mu}_i$, *i*=1,…,*C*, respectively. After running the *k*-Means algorithm on this data set, will you get the mean vectors $\boldsymbol{\mu}_i$ as the outcome?

Question 6. (**Problem 50 of Chapter 10**)



Section 10.14

50. Consider the use of multidimensional scaling for representing the points $\mathbf{x}_1 = (1, 0)^t$, $\mathbf{x}_2 = (0, 0)^t$, and $\mathbf{x}_3 = (0, 1)^t$ in one dimensions. To obtain a unique solution, assume that the image points satisfy $0 = y_1 < y_2 < y_3$.

(a) Show that the criterion function J_{ee} is minimized by the configuration with $y_2 = (1 + \sqrt{2})/3$ and $y_3 = 2y_2$.

(b) Show that the criterion function J_{ff} is minimized by the configuration with $y_2 = (2 + \sqrt{2})/4$ and $y_3 = 2y_2$.