Homework 8
Due: Friday, December 5, 11:59pm

Important: Submit the code on Blackboard and anything else on Gradescope.

1 Maximum Likelihood Estimation

Problem 7.1 In this question, we will see how to write the likelihood function for a time-series problem.

(a) (5 points) Suppose you are running a restaurant and you have collected daily revenues of your restaurant $x_1, x_2, \cdots, x_n$ ($n > 30$, $x_i \in \mathbb{R}$). You assume that the revenues for the first 30 days are independent and can be modeled by a Gaussian: $x_t \sim N(\mu, \sigma^2)$, $\forall 1 \leq t \leq 30$ where μ and σ^2 have to be inferred from the data. However, from the 31$^{\text{st}}$ day, you assume that you can predict the value of tomorrow's revenue based on the revenues of today, yesterday, and the day before yesterday. That is, $x_{t+1} = ax_t + bx_{t-1} + cx_{t-2} + \epsilon_{t+1}$, $\forall t \geq 30$, where ϵ_{t+1} is i.i.d. Gaussian noise $\epsilon_{t+1} \sim N(0, \lambda^2)$, and λ is assumed to be known and a, b, c must be learned from the data. Write the log likelihood function you should maximize to find a, b, c, μ, σ based on the revenues $x_1, \ldots, x_n$.

Hint: Remember that in the generative world, the data is coming from a distribution that belongs to the family of probability distributions you assume. So, to write the log likelihood, you first have to derive the pdf of the probability distribution on each training sample. Also, remember that the pdf of a Gaussian with mean μ and variance σ^2 is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

(b) (3 points) If you wrote the correct likelihood function, you should see that the likelihood depends separately on (a, b, c) and on (μ, σ). This implies that we can optimize the likelihood separately on the two groups of variables. Also, the part of the log likelihood that depends on (a, b, c) should have a familiar face. So, write the closed-form solution to compute a, b, c from the samples $x_1, \ldots, x_n$.

(c) (2 points) Suppose that the assumptions you made are wrong, how do you expect your trained model to behave on test data?

(d) (2 points) Write the log likelihood for the complete data (i.e., assuming x_{35} was observed, you get lazy and do not record the revenue). You may treat $x_{31}, \ldots, x_n$ as being generated by the autoregressive model.

(e) (2 points) Describe how you would use the EM algorithm to estimate parameters (a, b, c) in this setting. Clearly specify what is treated as missing, and what the E-step and M-step involve.

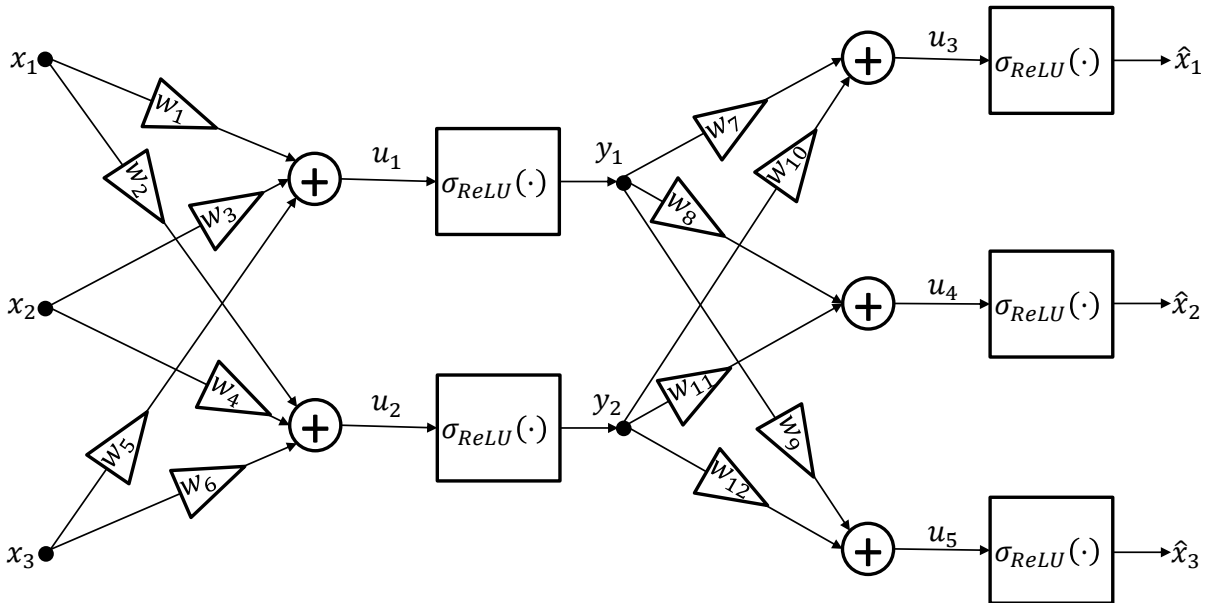
Problem 2: EM for disease diagnosis

Let $X = X_1, \ldots, X_n$ be random variables denoting the different symptoms and Y be the random variable denoting the diagnosis. Suppose the random variable Y takes on K possible values while each symptom

can take on L different values, namely, $Y \in \{y_1, y_2, \ldots, y_K\}$ and $X_j \in \{x_{1j}, x_{2j}, \ldots, x_{Lj}\}$. The probability model is parameterized by $\alpha_k = Prob(Y = y_k)$, $k = 1, 2, \ldots, K$ and $\beta_{i\ell k} = Prob(X_i = x_{il} \mid Y = y_k)$. Our goal is to estimate α, β from data.

1. Write down the log-likelihood function for the case when you are given complete data for m data instances $(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}, y^{(i)})$, $i = 1, 2, \ldots, m$.
2. Compute the ML estimates for using the log-likelihood function you wrote down in (1).
3. Assume now that for $k \leq i \leq m$ the diagnosis label y_i is missing or not observed. Compute the marginal log-likelihood of the data for this case.
4. E-step: Compute $c_k^{(i)} = P\{Y = y_k \mid x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}\}$ for the missing observations.
5. Now using the previous step write down the complete data likelihood function $\ell(\theta \mid \theta_t)$ described in class.
6. M-step: Optimize over the parameters to obtain estimates for the parameters. Write down the EM algorithm.

Problem 3: Auto-Encoder An autoencoder with ReLU activation function $\sigma_{ReLU}(t) = \max\{0, t\}$ is shown in the figure below. Let $\ell(\hat{x}_1, \hat{x}_2, \hat{x}_3, x_1, x_2, x_3) = \sum_{i=1}^{3}(\hat{x}_i - x_i)^2$ be the loss function, $(x_1 = 1, x_2 = -1, x_3 = 1)$ be a training sample, and $w_4 = w_7 = w_{11} = -1$ with the remaining weights equal to $+1$ be the initial weights. Compute the values of the partial derivatives of the loss with respect to: u_3, u_4, u_5,



$w_7, w_8, w_9, w_{10}, w_{11}, w_{12}$, and y_1, y_2 in the first backward pass iteration of the backpropagation algorithm. Re-sketch the network and indicate the partial derivatives in red color next to the corresponding variables.

Problem 4: Neural Networks Let $h(x) = \max\{2, |x| + 1\}$ and $\sigma_{ReLU}(t) = \max\{0, t\}$ be the Rectifier Linear Unit activation function. Find values of (β_1, γ_1) and (β_2, γ_2) such that for all x,

$$h(x) = \underbrace{\sigma_{ReLU}\left(\beta_1 + (\gamma_1 \cdot x)\right)}_{h_1(x)} + \underbrace{\sigma_{ReLU}\left(\beta_2 + (\gamma_2 \cdot x)\right)}_{h_2(x)}.$$

Sketch the graphs of $h_1(x), h_2(x)$, and $h(x)$ and properly label axes and key points.