CS760 Midterm Practice Questions

October 23, 2021

1 True/False

- Unsupervised learning methods deal with instances without labels, and can reveal patterns of the data.
- In cross validation, we train the classifier using all of the data, and predict the classification of the left-out set.
- High capacity models are more likely to overfit.
- A fully connected feedforward neural network has input $\mathbb{R}^2$, a first hidden ReLU layer with 4 units, a second hidden ReLU layer with 3 units, and a single sigmoid output unit. The number of parameters in the neural network (including offset weights) = 30.

2 Neural Networks

(a) Derive the derivative of ReLU activation function function $ReLU(x) = max(0, x)$. (5 pts)

(b) Derive the derivative of hyperbolic tangent activation function $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. (5 pts)

(c) Compare sigmoid, hyperbolic tangent, and ReLU activation functions. (5 pts)

3 Evaluation metrics

Suppose you trained a classifier for a spam filter. The prediction result on the test set is summarized in the following table. Here, "+" represents spam, and "-" means not spam.

Confidence positive	Correct class
0.95	+
0.85	+
0.8	-
0.7	+
0.55	+
0.45	-
0.4	+
0.3	+
0.2	-
0.1	-

Set the threshold for spam as 0.5 (i.e. if confidence positive is more than 0.5, it is classified as +). Calculate the following evaluation metrics.

- Accuracy (2.5 pts)
- False positive rate (2.5 pts)
- Precision (2.5 pts)
- Recall (2.5 pts)

Problem 4

Suppose you have n samples drawn $\{x_i\}_{i=1}^{n}$ i.i.d from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. What is the negative log-likelihood (NLL) of these samples? Derive the MLE (Maximum Likelihood Estimator) for μ and σ using these samples. Comment on the behaviour of your estimators when $n \to \infty$. What can you say about the convexity or strong convexity of the NLL function.

Problem 5

Let's say you have 3 points from a 3 dimensional space, namely $\mathbf{x}_1 = [2, -1, 0], \mathbf{x}_2 = [-1, 1, 1], \mathbf{x}_3 = [0, 1, 0]$. Also let, $y_1 = 0, y_2 = 5, y_3 = 2$. Assume there is some true $\mathbf{w} \in \mathbb{R}^3$ such that $\mathbf{w}^T\mathbf{x}_i = y_i$.

1. Write expression for closed form solution for $\mathbf{w}$. Does it exist? If yes, calculate it.
2. Do at least 2 steps of gradient descent (manually). Use learning rate 0.02 and initialize $\mathbf{w}$ to $[0, 1, 0]$. At each step compare the current estimate by gradient descent to the closed form solution.

Problem 6

Your friend has a biased coin i.e. probability of head is $\theta \neq 1/2$. You agreed to play a game with your friend. Your friend tosses the coin and you need to guess whats on the outcome of the toss. You are overly confident about your predictions so you agreed to the conditions that you will pay \$1 if you are wrong otherwise nobody pays anything.

1. Without knowing θ, what will be your strategy? What will be the expected amount of money you will be paying to your friend.
2. Can you estimate θ while playing the game? What will be the most likely estimate of θ after n rounds of play. What happens to your estimate if you keep playing the game forever. (You might go bankrupt though)
3. Suppose your friend's friend told you the θ, what will be your strategy? If you use this strategy what will be your expected loss?
4. By now may be you know you are loosing a lot of money without knowing θ. Instead of telling the true θ, your friend's friend told you that the true θ follows a Beta distribution with parameters α, β. How will you use this prior knowledge to improve your estimation? Derive the estimator using this prior knowledge and n rounds of play.