Theory Assignment 2 (Practice)

COMP 451 - Fundamentals of Machine Learning

Prof. William L. Hamilton

Winter 2021

Question 1 [7 points]

Recall that the logistic function is defined as

$$\sigma(z) = \frac{1}{1+e^{-z}}. \tag{1}$$

Also, note that the tanh function is defined as

$$\tanh(a) = \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}}. \tag{2}$$

Part 1 [3 points] Show that the logistic function and the tanh function are related by the following expression

$$\tanh(a) = 2\sigma(2a) - 1. \tag{3}$$

Part 2 [4 points] Consider a general linear combination of logistic sigmoid functions as follows:

$$f_{\mathbf{w}}(\mathbf{x}) = b + \sum_{j=0}^{m-1} \mathbf{w}[j]\sigma(\mathbf{x}[j]), \tag{4}$$

where $\mathbf{w}$ is a vector of weights and b is an intercept term. Show that this expression is equivalent to a linear combination of tanh functions of the following form:

$$h_{\mathbf{u}}(\mathbf{x}) = c + \sum_{j=0}^{m-1} \mathbf{u}[j] \tanh\left(\frac{\mathbf{x}[j]}{2}\right), \tag{5}$$

with weight vector $\mathbf{u}$ and intercept c. Your answer should show how $\mathbf{w}$ and b can be derived from $\mathbf{u}$ and c.

Question 2 [8 points]

Consider a linear model of the form

$$f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} = \sum_{j=0}^{m-1} \mathbf{w}[j]\mathbf{x}[j] \tag{6}$$

with a mean-squared empirical risk

$$R(\mathbf{w}) = \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{trn}}} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2. \tag{7}$$

Now, suppose that we add random Gaussian noise to the input feature vector. In particular, assume that the feature vector for each datapoint has the form

$$\mathbf{x}_i[j] = \tilde{\mathbf{x}}_i[j] + \epsilon_j, \tag{8}$$

where $\epsilon_j \sim \mathcal{N}(0, \sigma_i)$ is a normally distributed noise term with zero mean and $\tilde{\mathbf{x}}_i[j]$ denotes the original (un-noised) feature input. Show that minimizing the expected risk $\mathbb{E}[R(\mathbf{w})]$ under this noise distribution is equivalent to adding L2 regularization to a linear regression model with the original un-noised features $\tilde{\mathbf{x}}$. To show this, you should assume that the noise for the different feature dimensions are independent, which means that

$$\mathbb{E}[\epsilon_j \epsilon_k] = \begin{cases} \sigma_j^2 & \text{if } j = k \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

However, you can assume that variance of the noise is the same constant σ for all the ϵ_j, i.e., $\sigma_0 = \sigma_1 = ... = \sigma_{m-1} = \sigma$. In other words, you should assume that all the ϵ_j noise terms are *independent Gaussian variables* but that they have the same constant variance σ.

Question 3 [3 points]

Explain the conditions under which mini-batch gradient descent will be asymptotically faster than the closed-form solution for linear regression. You should consider asymptotic (i.e., big-$\mathcal{O}$) time complexity in your answer.

Question 4 [short answers; 2 points each]

Answer each question with 1-3 sentences for justification, potentially with equations/examples for support.

a) Suppose model A and B are both regression models trained using empirical risk minimization on the same dataset using using mean-squared error. Is the following statement true or false: If model A and B have equal statistical variance but model B has higher statistical bias, then model A will always have lower risk on the training dataset.

b) Suppose your closed form linear regression gives a singular matrix error. Describe two things you could do to address this issue.

Which of the following statements is false:

1. Models that underfit typically have higher statistical bias (and lower variance).
2. Gradient descent is guaranteed to converge to a local minimum of a function (but not necessarily the global minimum), as long as the step-size is small enough and the function is smooth.
3. L1 regularization is an effective way to enforce sparsity on the learned model parameters.
4. Knowing the Hessian matrix at every point is sufficient to test whether a function is convex.