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.

Solution. *We have that*

$$\begin{aligned}
2\sigma(2a) - 1 &= \frac{2e^{2a}}{1 + e^{2a}} - 1 && (6)\\
&= \frac{2e^{2a} - 1 - e^{2a}}{1 + e^{2a}} && (7)\\
&= \frac{e^{2a} - 1}{1 + e^{2a}} && (8)\\
&= \frac{e^a}{e^a}\frac{e^a - e^{-a}}{e^a + e^{-a}} && (9)\\
&= \frac{e^a - e^{-a}}{e^a + e^{-a}} && (10)
\end{aligned}$$

Now, for the second part of the question, we have that

$$h_{\mathbf{u}}(\mathbf{x}) = c + \sum_{j=0}^{m-1} \mathbf{u}[j] \tanh\left(\frac{\mathbf{x}[j]}{2}\right) \tag{11}$$
$$= c + \sum_{j=0}^{m-1} \mathbf{u}[j](2\sigma(\mathbf{x}[j]) - 1) \tag{12}$$
$$= c + \sum_{j=0}^{m-1} 2\mathbf{u}[j]\sigma(\mathbf{x}[j]) - \mathbf{u}[j] \tag{13}$$
$$= \left(c - \sum_{j=0}^{m-1} \mathbf{u}[j]\right) + \sum_{j=0}^{m-1} 2\mathbf{u}[j]\sigma(\mathbf{x}[j]), \tag{14}$$

which gives that

$$b = \left(c - \sum_{j=0}^{m-1} \mathbf{u}[j]\right), \tag{15}$$

and

$$\mathbf{w}[j] = 2\mathbf{u}[j], \forall j = 0..., m-1 \tag{16}$$

Question 2 [8 points]

Consider a linear model of the form

$$f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} = \sum_{j=0}^{m-1} \mathbf{w}[j]\mathbf{x}[j] \tag{17}$$

with a mean-squared empirical risk

$$R(\mathbf{w}) = \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{trn}}} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2. \tag{18}$$

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

$$\mathbf{x}_i[j] = \tilde{\mathbf{x}}_i[j] + \epsilon_j, \tag{19}$$

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

$$\mathbb{E}[\epsilon_j \epsilon_k] = \begin{cases} \sigma_j^2 & \text{if } j = k \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

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.

Solution.
Let $\boldsymbol{\epsilon} = [\epsilon_0, \epsilon_1, ..., \epsilon_{m-1}]^\top$ *denote a vector with the noise values for each feature dimension. Note that we have that* $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$. *Furthermore, in our derivation, we will use the fact that for any* j *and* k *where* $j \neq k$ *we have that*

$$\mathbb{E}[f(\boldsymbol{\epsilon}) + C_1\epsilon_j + C_2\epsilon_j\epsilon_k] = \mathbb{E}[f(\boldsymbol{\epsilon})] + C_1\mathbb{E}[\epsilon_j] + C_2\mathbb{E}[\epsilon_j\epsilon_k] \tag{21}$$
$$= \mathbb{E}[f(\boldsymbol{\epsilon})]. \tag{22}$$

where $f(\boldsymbol{\epsilon})$ is an arbitrary function of the noise and C_1 and C_2 are constants that do not depend on the noise. Now, we have that

$$
\begin{aligned}
\mathbb{E}[R(\mathbf{w})] &= \mathbb{E}\left[\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{trn}} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2\right] && (23)\\
&= \mathbb{E}\left[\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{trn}} y_i^2 - 2y_i\mathbf{w}^\top \mathbf{x}_i + (\mathbf{w}^\top \mathbf{x}_i)^2\right] && (24)\\
&= \mathbb{E}\left[\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{trn}} y_i^2 - 2y_i\mathbf{w}^\top (\tilde{\mathbf{x}}_i + \boldsymbol{\epsilon}) + (\mathbf{w}^\top \mathbf{x}_i)^2\right] && (25)\\
&= \mathbb{E}\left[\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{trn}} y_i^2 - 2y_i\mathbf{w}^\top \tilde{\mathbf{x}}_i + (\mathbf{w}^\top \mathbf{x}_i)^2\right] - \sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{trn}} 2y_i\mathbf{w}^\top \mathbb{E}[\boldsymbol{\epsilon}] && (26)\\
&= \mathbb{E}\left[\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{trn}} y_i^2 - 2y_i\mathbf{w}^\top \tilde{\mathbf{x}}_i + (\mathbf{w}^\top (\tilde{\mathbf{x}}_i + \boldsymbol{\epsilon}))^2\right] && (27)\\
&= \mathbb{E}\left[\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{trn}} y_i^2 - 2y_i\mathbf{w}^\top \tilde{\mathbf{x}}_i + (\sum_{j=0}^{m-1} \mathbf{w}[j]\tilde{\mathbf{x}}[j] + \mathbf{w}[j]\epsilon_j)^2\right] && (28)\\
&= \mathbb{E}\Bigg[\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{trn}} y_i^2 - 2y_i\mathbf{w}^\top \tilde{\mathbf{x}}_i + \Bigg(\sum_{j=0}^{m-1} \mathbf{w}[j]^2\tilde{\mathbf{x}}[j]^2 && (29)\\
&+ \mathbf{w}[j]^2\epsilon_j^2 + \mathbf{w}[j]^2\tilde{\mathbf{x}}[j]\epsilon_j + \sum_{k=0,k\neq j}^{m-1} \mathbf{w}[j]\mathbf{w}[k]\tilde{\mathbf{x}}[j]\tilde{\mathbf{x}}[k] + \mathbf{w}[j]\mathbf{w}[k]\epsilon_j\epsilon_k + \mathbf{w}[j]\tilde{\mathbf{x}}[j]\mathbf{w}[k]\epsilon_k\Bigg)\Bigg] && (30)\\
&= \mathbb{E}\left[\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{trn}} y_i^2 - 2y_i\mathbf{w}^\top \tilde{\mathbf{x}}_i + \left(\sum_{j=0}^{m-1} \mathbf{w}[j]^2\tilde{\mathbf{x}}[j]^2 + \mathbf{w}[j]^2\epsilon_j^2 + \sum_{k=0,k\neq j}^{m-1} \mathbf{w}[j]\mathbf{w}[k]\tilde{\mathbf{x}}[j]\tilde{\mathbf{x}}[k]\right)\right] && (31)\\
&= \mathbb{E}\left[\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{trn}} y_i^2 - 2y_i\mathbf{w}^\top \tilde{\mathbf{x}}_i + \left(\sum_{j=0}^{m-1} \mathbf{w}[j]^2\tilde{\tilde{\mathbf{x}}}[j]^2 + \sum_{k=0,k\neq j}^{m-1} \mathbf{w}[j]\mathbf{w}[k]\tilde{\mathbf{x}}[j]\tilde{\mathbf{x}}[k]\right)\right] + \mathbb{E}\left[\sum_{j=0}^{m-1} \mathbf{w}[j]^2\epsilon_j^2\right] && (32)\\
&= \mathbb{E}\left[\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{trn}} y_i^2 - 2y_i\mathbf{w}^\top \tilde{\mathbf{x}}_i + (\mathbf{w}^\top \tilde{\mathbf{x}}_i)^2\right] + \sum_{j=0}^{m-1} \mathbf{w}[j]^2\mathbb{E}[\epsilon_j^2] && (33)\\
&= \mathbb{E}\left[(y_i - \mathbf{w}^\top \tilde{\mathbf{x}})^2\right] + \sigma^2\|\mathbf{w}\|^2, && (34)
\end{aligned}
$$

which corresponds to L2 regularization with the strength of the regularization given by the variance term σ^2.

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.

Solution.

The cost of each iteration of mini-batch gradient descent is $\mathcal{O}(mB^2)$, assuming a batch-size of B. Thus, if it takes K iterations for the mini-batch gradient descent to converge, we have that the overall time complexity of $\mathcal{O}(kB^2m)$

In contrast, the closed form solution has time complexity $\mathcal{O}(n^2 + m^3)$.

Thus, we can expect the mini-batch solution to be faster when

$$KmB^2 < n^2 + m^3$$

Now, we can assume that B grows sub-linearly in n (since otherwise we would end up with full-batch gradient descent) and we similarly can assume that m grows sub-linearly in n (since otherwise we would end up with an *ill-formed/singular regression problem). Thus, we can expect gradient descent will be asymptotically faster whenever $K < n^2$. (Note there are other reasonable answers here; as long as reasonable assumptions are made.)*

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.

Solution.

a) *Yes, the training mean-squared error (i.e., risk) is equal to the sum of the statistical variance and bias. Thus, if they have equal statistical variance but B has higher bias, the sum (i.e., the error) must be larger for model B.*

b) *Add L2-regularization or collect more training data.*

c) *The second statement is false. Gradient descent might converge to a saddle point.*