Homework #2

Claire Jung

May 1, 2023

Problem 1

First, we know from class that we can rewrite the conditional risk for $R(h^*|x)$ as:

$$R(h^*|x) = \sum_{h(x)\neq c}^{C} p(y = c|x)$$

Now, computing the conditional probability of the randomized classifier:

$$R(h_r|x) = \sum_{c=1}^{C}\sum_{c'=1}^{C} L_{0/1}(c', c)q_r(c_r = c|x)p(y = c|x)$$

The zero one loss will pick out terms from the sum, and the sum of the random probabilities must be one, so we can rewrite the classifier as:

$$R(h_r|x) = \sum_{c=1}^{C} p(y = c|x)(1 - q_r(c_r = c|x))$$

Now, we need to show that for any q the inequality holds, so lets first look at the minimum of our randomized classifier risk. In order to minimize the risk, we want the largest probability to not contribute, which is the probability of the correctly classified example. In other words, the risk is at a minimum when $q_r(c_r = h(x)|x) \to 1$ for the correct c, and all other $q_r(c_r \neq h(x)|x) \to 0$. Therefore,

$$R(h_r|x)_{min} = \sum_{c\neq h(x)}^{C} p(y = c|x) = R(h^*|x)$$

Clearly, the conditional risk for the randomized classifier can always be larger for any function q which weighs any classification that is not correct. If we choose $q_r(c_r = c) \to 1$ for c that has minimizes$p(y = c|x)$,

$$R(h_r|x)_{max} = \sum_{c\neq c_{min}}^{C} p(y = c|x) + p(y = h(x)|x) > R(h^*|x)$$

Therefore,

$$R(h_r|x) \geq R(h^*|x)$$

Problem 2

For L_2 regularized regression we can write our weights as:

$$w^* = (X^T X)^{-1} X^T Y$$

and the weights for the augmented data as:

$$w^* = (X'^T X' + \lambda I)^{-1} X'^T Y'$$

Clearly we want:

$$(X^T X) = (X'^T X' + \lambda I) \text{ and } X^T Y = X'^T Y'$$

This can be done by simply adding more terms to the original data, to create the augmented data:

$$X' = \begin{pmatrix} X \\ \sqrt{\lambda} I \end{pmatrix} \text{ and } Y = \begin{pmatrix} Y \\ 0 \end{pmatrix}$$

This satisfies the conditions above.

More concretely,

$$\sum_{i=1}^{N} (y_i - x_i^T w)^2 + \sum_{j=1}^{d} w_j^2$$

$$\sum_{i=1}^{N} (y_i - x_i^T w)^2 + \sum_{j=1}^{d} (0 \times y_j - \sqrt{\lambda} w_j)^2$$

And the additional terms are just tacked on.

Problem 3

Softmax takes the form:

$$\hat{p}(y = c_i | x; w) = \frac{\exp(w_i \cdot x)}{\sum_j^C \exp(w_j \cdot x)}$$

Taking the log odds between two classes,

$$\ln\left(\frac{\hat{p_1}}{\hat{p_2}}\right) = \ln\left(\frac{\exp(w_1 \cdot x)}{\exp(w_2 \cdot x)}\right) = (w_1 - w_2) \cdot x$$

is a linear function. In the binary case:

$$\frac{\exp(w_1 \cdot x)}{\exp(w_1 \cdot x) + \exp(w_2 \cdot x)} \times \frac{\exp(-w_1 \cdot x)}{\exp(-w_1 \cdot x)} = \frac{1}{1 + \exp((w_2 - w_1) \cdot x)} = \sigma(v \cdot x)$$

where $v = (w_2 - w_1)$.

Problem 4

The log loss for a single data point with a single class is,

$$LL = -y \ln \hat{p}(y|x; w, b)$$

Where we use softmax probability. Over multiple data samples,

$$LL = -1/N \sum_{i=1}^{N} y_i \ln \hat{p}(y|x_i; w)$$

And now, over multiple classes, with regularization

$$LL = -1/N \sum_{i=1}^{N} \sum_{j=1}^{C} t_{i,j} \ln \hat{p}(y = c|x_i; w_j) + \lambda ||W||^2$$

Where $t_{i,j}$ is the one-hot encoding of the labels. Taking a derivative of the data with respect to weights for a single class,

$$\frac{\partial LL}{\partial w_q} = -1/N \frac{\partial}{\partial w_q} \sum_{i=1}^{N} \sum_{j=1}^{C} t_{i,j} [w_j \cdot x_i + b_j - \ln(\sum_{k=1}^{C} \exp(w_k \cdot x_i + b_k))]$$

The log term will return our original probability, as the derivative picks out one term in the sum returning $x_i \exp(w_q \cdot x_i + b_q)$ and the log will keep the sum in the denominator. We get an additional $t_{i,q}$ factor also by picking out just one non-zero term.

$$\frac{\partial LL}{\partial w_q} = -1/N \sum_{i=1}^{N} \sum_{j=1}^{C} t_{i,j} [t_{i,j=q} x_i - x_i \hat{p}(y = q|x_i; w_q, b_q)]$$

We can now remove the sum, as it only picks out a terms where $j = q$.

$$\frac{\partial LL}{\partial w_q} = -1/N \sum_{i=1}^{N} x_i [t_{i,q} - \hat{p}(y = q|x_i; w_q, b_q)]$$

The regularization term is simply $2\lambda w_q$

Similarly for b,

$$\frac{\partial LL}{\partial b_q} = -1/N \sum_{i=1}^{N} [t_{i,q} - \hat{p}(y = q|x_i; w_q, b_q)]$$

In the stochastic setting, we are computing the loss for a single example. Therefore our gradients are:

$$\left(\frac{\partial LL}{\partial w_q}\right)_i = -x_i [t_{i,q} - \hat{p}(y = q|x_i; w_q, b_q)] + 2\lambda w_q$$

$$\left(\frac{\partial LL}{\partial b_q}\right)_i = -[t_{i,q} - \hat{p}(y = q|x_i; w_q, b_q)]$$

And our update equations,

$$w = w + \eta(x_i[t_{i,j} - \hat{p}(y|x_i; w, b)] - 2\lambda w)$$

$$b = b + \eta[t_{i,j} - \hat{p}(y|x_i; w, b)]$$

Where, $t_{i,j}$ and $\hat{p}(y|x_i; w, b)$ are matrices computed over all classes.

Problem 5

See attached notebook.