Linear Models

Exercise 1 : Derivative of $l_\sigma(c, y(\mathbf{x}))$

$$\text{Show that } \nabla l_\sigma(c, y(\mathbf{x})) = -\delta \cdot \mathbf{x},$$
$$\text{i.e., } \frac{\partial}{\partial w_i}[l_\sigma(c, y(\mathbf{x}))] = -\delta \cdot x_i$$

with $\delta = c - y(\mathbf{x})$ for $y(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x})$.

Answer

$$\begin{aligned} l_\sigma(c, y(\mathbf{x})) &= -c \cdot \log(y(\mathbf{x})) - (1-c) \cdot \log(1 - y(\mathbf{x})) \\ &= -c \cdot \log(\sigma(\mathbf{w}^T\mathbf{x})) - (1-c) \cdot \log(1 - \sigma(\mathbf{w}^T\mathbf{x})) \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial w_i}[l_\sigma(c, y(\mathbf{x}))] &= -c \cdot \frac{1}{\sigma(\mathbf{w}^T\mathbf{x})} \cdot (\sigma(\mathbf{w}^T\mathbf{x})(1 - \sigma(\mathbf{w}^T\mathbf{x}))) \cdot x_i \\ &\quad - (1-c) \cdot \frac{1}{1 - \sigma(\mathbf{w}^T\mathbf{x})} \cdot (-1) \cdot (\sigma(\mathbf{w}^T\mathbf{x})(1 - \sigma(\mathbf{w}^T\mathbf{x}))) \cdot x_i \\ &= -c \cdot (1 - \sigma(\mathbf{w}^T\mathbf{x})) \cdot x_i \\ &\quad - (1-c) \cdot (-1) \cdot \sigma(\mathbf{w}^T\mathbf{x}) \cdot x_i \\ &= -c \cdot x_i + c \cdot \sigma(\mathbf{w}^T\mathbf{x}) \cdot x_i + \sigma(\mathbf{w}^T\mathbf{x}) \cdot x_i - c \cdot \sigma(\mathbf{w}^T\mathbf{x}) \cdot x_i \\ &= -c \cdot x_i + \sigma(\mathbf{w}^T\mathbf{x}) \cdot x_i \\ &= -c \cdot x_i + y(\mathbf{x}) \cdot x_i \\ &= -\delta \cdot x_i \end{aligned}$$

Exercise 2 : Batch Gradient Descent

(a) Explain the difference of LMS and BGD that allows to overcome a significant flaw of the former.

Answer

The values of $\mathbf{\Delta w}$ are summed up to allow an update step respecting each $(\mathbf{x}, c) \in D$ from the same $\mathbf{w}$.

(b) How is that related to the idea of a global loss?

Answer

The global loss $L(\mathbf{w})$ considers each $(\mathbf{x}, c) \in D$ at the same $\mathbf{w}$.

(c) Verify that, indeed, $\nabla L(\mathbf{w}) = \sum_{(\mathbf{x},c)\in D} \nabla l(c, y(\mathbf{x}))$.

Answer

This holds because of the linearity of differentiation, applied on the definition of

$$L(\mathbf{w}) = \sum_{(\mathbf{x},c)\in D} l(c, y(\mathbf{x})).$$

(d) Why is that needed to justify the BGD algorithm?

Answer

By summing up the $\mathbf{\Delta w}$ for each $(\mathbf{x}, c) \in D$, BGD sum the derivatives of the pointwise loss $l(c, y(\mathbf{x}))$. In the end, this sum must form the derivative of the global loss $L(\mathbf{w})$.

Exercise 3 : Convergence criterion

In algorithms like LMS and BGD, we analyze the global loss $L(\mathbf{w}_t)$ or the norm of the loss gradient $||\nabla L(\mathbf{w}_t)||$ at the time step t.

In this exercise, we will compare three choices of comparison values in the convergence criterion and see that they are approximately equivalent.

(a) By using the update rule $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \nabla L(\mathbf{w}_t)$, express $\Delta L = L(\mathbf{w}_t) - L(\mathbf{w}_{t+1})$ in terms of $||\nabla L(\mathbf{w}_t)||$.

Hint: You may estimate using the Taylor formula: $f(x) \approx f(a) + \nabla f(a)^T (x - a)$ for cleverly chosen f, x and a.

Answer

With $f := L$, $x := \mathbf{w}_t - \eta \cdot \nabla L(\mathbf{w}_t)$ and $a := \mathbf{w}_t$, we get:

$$\begin{aligned} L(\mathbf{w}_t - \eta \cdot \nabla L(\mathbf{w}_t)) \approx & L(\mathbf{w}_t) + (\nabla L(\mathbf{w}_t))^T (\mathbf{w}_t - \eta \cdot \nabla L(\mathbf{w}_t) - \mathbf{w}_t) \\ = & L(\mathbf{w}_t) - \eta \cdot (\nabla L(\mathbf{w}_t))^T (\nabla L(\mathbf{w}_t)) \end{aligned}$$

and thus

$$\Delta L(\mathbf{w}) \approx \eta \cdot ||\nabla L(\mathbf{w}_t)||^2.$$

(b) Express $||\Delta \mathbf{w}|| = ||\mathbf{w}_{t+1} - \mathbf{w}_t||$ in terms of $||\nabla L(\mathbf{w}_t)||$.

Answer

$$||\Delta \mathbf{w}|| = || - \eta \cdot \nabla L(\mathbf{w}_t)|| = \eta \cdot ||\nabla L(\mathbf{w}_t)||$$

(c) What does that mean for the choice of ε in the convergence criterion? Why would we prefer to use $||\nabla L(\mathbf{w}_t)||$?

Answer

For $||\Delta \mathbf{w}||$, one would have to choose a significantly smaller ε than for $||\nabla L(\mathbf{w}_t)||$. For ΔL, this would have to be even smaller, which might be a problem when representing numbers as float values.

Exercise 4 : Overfitting and train-test leakage

(a) What is the experimental setup of choice when trying to detect overfitting?

Answer

Evaluation on an annotated validation set.

(b) What are methods to mitigate overfitting?

Answer

Increasing quantity and/or quality of the training data, early stopping, regularization.

(c) What must be payed attention to when performing a train-validation split on the following datasets in the given problems?

(c1) Detecting pneumonia from chest x-rays. Data includes 112,120 unique images from 30,805 unique patients.

Answer

While having the label classes represented each in training and validation set, the images of one individual patient must not be split up. Otherwise, the model will learn to classify that particular patient and not generalize among patients.

This is a classic mistake that was famously made in Andrew Ng's research group.

(c2) Given 1000 voice recordings (single sentences) of 100 people in total from 5 German cities. The model should be able to classify the dialects of arbitrary people into one of these cities.

Answer

Don't split up recordings of an individual person. Split up each of the cities.

(c3) Given 1000 voice recordings (single sentences) of 100 people in total from 5 German cities. The model should be able to rate the dialects of arbitrary people from all over Germany by intelligibility.

Answer

Don't split up recordings of an individual person. Don't split up recordings from each of the cities, rather keep the recordings from individual cities together. Otherwise, there won't be generalization among cities.

(c4) Given 1000 voice recordings (single sentences) of 100 people in total from 5 German cities. The model should be able to classify the person that said a given sentence.

Answer

Split up the recordings of an individual person. This means also splitting up the recordings of an individual city.