Linear Models

Exercise 1 : Properties of the Sigmoid Function

This exercise regards some mathematical properties of the sigmoid function σ, which make it very suitable for machine learning.

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

(a) Show that $\sigma(-x) = 1 - \sigma(x)$.

Answer

Starting from right side is much easier. Add and multiply by 1 in form of e^x/e^x.
$1 - \sigma(x) = 1 - \frac{1}{1+e^{-x}} = \frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}} = \frac{e^{-x}}{1+e^{-x}} = \frac{e^{-x}}{1+e^{-x}} \cdot \frac{e^x}{e^x} = \frac{1}{1+e^x} = \sigma(-x)$

(b) Show that the derivative of the sigmoid function is $\frac{\partial\sigma(x)}{\partial x} = \sigma(x)(1 - \sigma(x))$.

Answer

This is best done by chain rule to the $.^{-1}$ notation and using the result from a)
$\frac{\partial\sigma(x)}{\partial x} = \frac{\partial}{\partial x}[(1+e^{-x})^{-1}] = (-1) \cdot (1+e^{-x})^{-2} \cdot e^{-x} \cdot (-1) = \frac{e^{-x}}{1+e^{-x}} \cdot \frac{1}{1+e^{-x}} =$
$\frac{e^x}{e^x} \cdot \frac{e^{-x}}{1+e^{-x}} \cdot \frac{1}{1+e^{-x}} = \sigma(-x)\sigma(x) = (1-\sigma(x))\sigma(x)$

Exercise 2 : Logistic Regression

For the task of binary sentiment classification on movie review texts, we represent each input text by the 6 features $x_1 ... x_6$ shown for three training examples together with the ground-truth class label ($0 =$negative, $1 =$positive) in the following table.

Feat.	Definition	Example 1	Example 2	Example 3
x_1	Count of positive lexicon terms	3	1	5
x_2	Count of negative lexicon terms	2	5	2
x_3	1 if “no“ in doc, 0 otherwise	1	0	1
x_4	Count of 1st and 2nd pronouns	3	4	4
x_5	1 if “!” in doc, 0 otherwise	1	1	0
x_6	Word count	$\ln(66) = 4.19$	$\ln(119) = 4.77$	$\ln(45) = 3.81$
c	Sentiment class	1	0	1

A logistic regression model is given as $y(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x})$ with

$$\mathbf{w} = [0.21, 1.58, -1.36, -1.17, -0.17, 2.0, 0.14]^T$$

.

(a) Calculate the class probabilites $P(\mathbf{C} = 1 \mid \mathbf{X} = \mathbf{x}; \mathbf{w})$ and $P(\mathbf{C} = 0 \mid \mathbf{X} = \mathbf{x}; \mathbf{w})$ for each example and the given weights.

Answer

Example 1:

$$\begin{aligned}
P(\mathbf{C}=1 \mid \mathbf{X}=\mathbf{x};\mathbf{w}) &= \sigma(\mathbf{w}^T\mathbf{x})\\
&= \sigma([0.21, 1.58, -1.36, -1.17, -0.17, 2.0, 0.14]\cdot[1, 3, 2, 1, 3, 1, 4.19]^T)\\
&= \sigma(3.1352)\\
&= 0.9583\\
P(\mathbf{C}=0 \mid \mathbf{X}=\mathbf{x};\mathbf{w}) &= 1-\sigma(\mathbf{w}^T\mathbf{x})\\
&= 1-0.9583\\
&= 0.0417
\end{aligned}$$

Example 2:

$$\begin{aligned}
P(\mathbf{C}=1 \mid \mathbf{X}=\mathbf{x};\mathbf{w}) &= \sigma(\mathbf{w}^T\mathbf{x})\\
&= \sigma([0.21, 1.58, -1.36, -1.17, -0.17, 2.0, 0.14]\cdot[1, 1, 5, 0, 4, 1, 4.77]^T)\\
&= \sigma(-3.0222)\\
&= 0.0464\\
P(\mathbf{C}=0 \mid \mathbf{X}=\mathbf{x};\mathbf{w}) &= 1-\sigma(\mathbf{w}^T\mathbf{x})\\
&= 1-0.0464\\
&= 0.9436
\end{aligned}$$

Example 3:

$$\begin{aligned}
P(\mathbf{C}=1 \mid \mathbf{X}=\mathbf{x};\mathbf{w}) &= \sigma(\mathbf{w}^T\mathbf{x})\\
&= \sigma([0.21, 1.58, -1.36, -1.17, -0.17, 2.0, 0.14]\cdot[1, 5, 2, 1, 4, 0, 3.81]^T)\\
&= \sigma(4.0734)\\
&= 0.9833\\
P(\mathbf{C}=0 \mid \mathbf{X}=\mathbf{x};\mathbf{w}) &= 1-\sigma(\mathbf{w}^T\mathbf{x})\\
&= 1-0.88\\
&= 0.0167
\end{aligned}$$

(b) Compute $\Delta\mathbf{w}$ for one iteration of the BGD algorithm with a learning rate of $\eta = 0.1$.

Answer

Remarks: $y(\mathbf{x})$ were already calculated in (a); the values for $\boldsymbol{\Delta}\mathbf{w}$ are written individually here, but would be summed directly in the BGD algorithm.

Example	$y(\mathbf{x})$	c	$\delta = c - y(\mathbf{x})$	$\boldsymbol{\Delta}\mathbf{w} = \eta\cdot\delta\cdot\mathbf{x}$
1	0.9583	1	0.0417	$[0.004, 0.013, 0.008, 0.004, 0.013, 0.004, 0.017]^T$
2	0.0464	0	-0.0464	$[-0.005, -0.005, -0.023, -0.0, -0.019, -0.005, -0.022]^T$
3	0.9833	1	0.0167	$[0.002, 0.008, 0.003, 0.002, 0.007, 0.0, 0.006]^T$
$\sum$				$[0.001, 0.016, -0.012, 0.006, 0.001, -0.001, 0.001]^T$

(c) For the updated weights $\mathbf{w}+\Delta\mathbf{w}$, calculate the class probabilites $P(\mathbf{C}=0 \mid \mathbf{X}=\mathbf{x};\mathbf{w}+\Delta\mathbf{w})$ and $P(\mathbf{C}=0 \mid \mathbf{X}=\mathbf{x};\mathbf{w}+\Delta\mathbf{w})$ for each example. Compare them to your solution in (a); what can you observe?

Answer

$$\begin{aligned}&\mathbf{w}+\Delta\mathbf{w}\\&=[0.21,1.58,-1.36,-1.17,-0.17,2.0,0.14]^T+[0.001,0.016,-0.012,0.006,0.001,-0.001,0.001]^T\\&=[0.211,1.596,-1.372,-1.164,-0.169,1.999,0.141]^T\end{aligned}$$

Example 1:

$$\begin{aligned}P(\mathbf{C}=1\mid\mathbf{X}=\mathbf{x};\mathbf{w})&=\sigma(\mathbf{w}^T\mathbf{x})\\&=\sigma([0.211,1.596,-1.372,-1.164,-0.169,1.999,0.141]\cdot[1,3,2,1,3,1,4.19]^T)\\&=\sigma(3.1724)\\&=0.9598\\P(\mathbf{C}=0\mid\mathbf{X}=\mathbf{x};\mathbf{w})&=1-\sigma(\mathbf{w}^T\mathbf{x})\\&=1-0.9583\\&=0.0402\end{aligned}$$

Example 2:

$$\begin{aligned}P(\mathbf{C}=1\mid\mathbf{X}=\mathbf{x};\mathbf{w})&=\sigma(\mathbf{w}^T\mathbf{x})\\&=\sigma([0.211,1.596,-1.372,-1.164,-0.169,1.999,0.141]\cdot[1,1,5,0,4,1,4.77]^T)\\&=\sigma(-3.0574)\\&=0.0449\\P(\mathbf{C}=0\mid\mathbf{X}=\mathbf{x};\mathbf{w})&=1-\sigma(\mathbf{w}^T\mathbf{x})\\&=1-0.0464\\&=0.9551\end{aligned}$$

Example 3:

$$\begin{aligned}P(\mathbf{C}=1\mid\mathbf{X}=\mathbf{x};\mathbf{w})&=\sigma(\mathbf{w}^T\mathbf{x})\\&=\sigma([0.211,1.596,-1.372,-1.164,-0.169,1.999,0.141]\cdot[1,5,2,1,4,0,3.81]^T)\\&=\sigma(4.1442)\\&=0.9844\\P(\mathbf{C}=0\mid\mathbf{X}=\mathbf{x};\mathbf{w})&=1-\sigma(\mathbf{w}^T\mathbf{x})\\&=1-0.88\\&=0.0145\end{aligned}$$

Comparison: the gradient descent step adjusted the weights in such the way that each predicted class moves (slightly) closer to the true label.

Exercise 3 : Regularization

Suppose we are estimating the regression coefficients in a linear regression model by minimizing the objective function $\mathcal{L}$.

$$\mathcal{L}(\mathbf{w})=\mathsf{RSS}_{tr}(\mathbf{w})+\lambda\mathbf{w}^T\mathbf{w}$$

The term $\mathsf{RSS}_{tr}(\mathbf{w})=\sum_{(\mathbf{x}_i,y_i)\in D_{tr}}\left(y_i-\mathbf{w}^T\mathbf{x}_i\right)^2$ refers to the residual sum of squares computed on the set D_{tr} that is used for parameter estimation. Assume that we can also compute an RSS_{test} on a separate set D_{test} that we don't use during training.

When we vary the hyperparameter λ, starting from 0 and gradually increase it, what will happen to the following quantities? Explain your answers.

(a) The value of $\mathsf{RSS}_{tr}(\mathbf{w})$ will...

☐ remain constant.
☒ steadily increase.
☐ steadily decrease.
☐ increase initially, then eventually start decreasing in an inverted U shape.
☐ decrease initially, then eventually start increasing in a U shape.

Answer

The increasing regularization term moves the minimum point of $\mathcal{L}$ to a parameter vector that fits the training data less well as measured by RSS alone. Hence the training residuals will only increase.

(b) The value of $\mathsf{RSS}_{test}(\mathbf{w})$ will...

☐ remain constant.
☐ steadily increase.
☐ steadily decrease.
☐ increase initially, then eventually start decreasing in an inverted U shape.
☒ decrease initially, then eventually start increasing in a U shape.

Answer

We initially remove the error due to overfitting, which has the potential to improve the fit on unseen data. As $\lambda \to \infty$, the norm of the learned parameters $||\mathbf{w}|| \to 0$, and the test residuals eventually increase again.