# Linear Models

Exercise 1 : Properties of the Sigmoid Function

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

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

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

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

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 | 4 | 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 | $\ln(66) = 4.19$ | $\ln(72) = 4.77$ | $\ln(45) = 3.81$ |
| $c$ | Sentiment class | 1 | 0 | 1 |

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

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

.

(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.

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

(c) Calculate the class probabilites $P(C = 1 \mid \mathbf{x}; \mathbf{w})$ and $P(C = 0 \mid \mathbf{x}; \mathbf{w})$ for each example and the updated weights $\hat{\mathbf{w}} = \mathbf{w} + \Delta \mathbf{w}$. Compare them to your solution in (a); what can you observe?

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 $\mathsf{RSS}_{tr}(\mathbf{w}) = \sum_{(x_i,y_i)\in D_{tr}} \left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2$ 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 $\mathsf{RSS}_{test}$ on a separate set $D_{test}$ that we don't use during training.

When we vary the hyperparameter $\lambda$, 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.

(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.