Machine Learning and Computational Statistics

Theodoros Georgiopoulos

Homework 8

Exercises for Unit 8: Logistic Regression classifier

Solution for exercise 1

We have a dataset $Y = \{(y_i, \boldsymbol{x_i}, i = 1, \dots, N)\}$ where $y_i \in \{0, 1\}$ is the class label for vector $\boldsymbol{x}_i \in R^l$. Lets extract the gradient descent logistic regression classifier. Because we have binary classification, the model is

$$\ln \frac{P(y=1|\boldsymbol{x})}{P(y=0|\boldsymbol{x})} = \boldsymbol{\theta}^T \boldsymbol{x}$$

which becomes

$$P(y=1|\boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{\theta}^T\boldsymbol{x}}} = \sigma(\boldsymbol{\theta}^T\boldsymbol{x})$$

using $P(y = 0|\boldsymbol{x}) = 1 - P(y = 1|\boldsymbol{x})$. The aim is to determine the $\boldsymbol{\theta}$ that maximizes $P(y_i = 1|\boldsymbol{x}_i)$ for all $\boldsymbol{x}_i$ with $y_i = 1$ and maximizes $P(y_i = 0|\boldsymbol{x}_i)$ for all $\boldsymbol{x}_i$ with $y_i = 0$. So, we want to maximize the likelihood

$$\prod_{\boldsymbol{x}_i : y_i = 1} P(y_i = 1|\boldsymbol{x}_i) \prod_{\boldsymbol{x}_i : y_i = 0} P(y_i = 0|\boldsymbol{x}_i).$$

We can make it simpler and maximize

$$\prod_{i=1}^{N} P(y_i = 1|\boldsymbol{x}_i)^{y_i} P(y_i = 0|\boldsymbol{x}_i)^{1-y_i} = \prod_{i=1}^{N} \sigma(\boldsymbol{\theta}^T\boldsymbol{x}_i)^{y_i}(1-\sigma(\boldsymbol{\theta}^T\boldsymbol{x}_i))^{1-y_i}.$$

Because we will use gradient descent, we want to mizimize the negative log likelihood, which is

$$\begin{aligned}
L(\boldsymbol{\theta}) &= -\sum_{i=1}^{N} \left(\sigma(\boldsymbol{\theta}^T\boldsymbol{x}_i)^{y_i} + (1-\sigma(\boldsymbol{\theta}^T\boldsymbol{x}_i))^{1-y_i}\right) \\
&= -\sum_{i=1}^{N} \left(y_i \ln(\sigma(\boldsymbol{\theta}^T\boldsymbol{x}_i)) + (1-y_i)\ln(1-\sigma(\boldsymbol{\theta}^T\boldsymbol{x}_i))\right).
\end{aligned}$$

Taking the gradient with respect to $\boldsymbol{\theta}$ and using $\frac{\partial \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i)}{\partial \boldsymbol{\theta}} = \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i)(1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i))$ gives

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) &= -\sum_{i=1}^{N} \left(\frac{y_i}{\sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i)} \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i)(1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i))\boldsymbol{x}_i - \frac{1 - y_i}{1 - \sigma(\boldsymbol{\theta}^{\boldsymbol{T}} \boldsymbol{x}_i)} \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i)(1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i))\boldsymbol{x}_i \right) \\
&= -\sum_{i=1}^{N} \left(y_i (1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i))\boldsymbol{x}_i - (1 - y_i)\sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i)\boldsymbol{x}_i \right) \\
&= -\sum_{i=1}^{N} \left(y_i (1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i)) - (1 - y_i)\sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i) \right) \boldsymbol{x}_i \\
&= -\sum_{i=1}^{N} (y_i - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i))\boldsymbol{x}_i \\
&= \boldsymbol{X}^T (\boldsymbol{\sigma}(\boldsymbol{\theta}^T \boldsymbol{x}) - \boldsymbol{y})
\end{aligned}$$

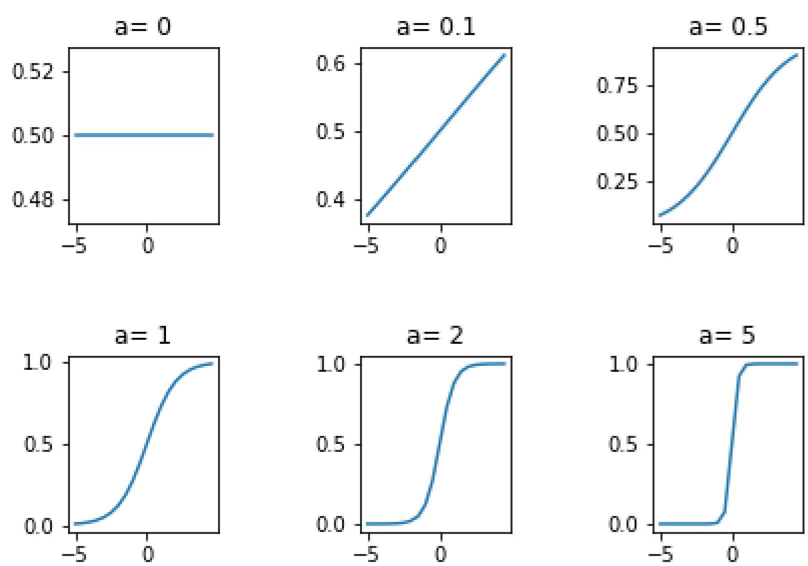
where $\boldsymbol{X}^T = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x}_N]$, $\boldsymbol{y} = [y_1, y_2, \ldots, y_N], \boldsymbol{\sigma}(\boldsymbol{\theta^T x}) = [\sigma(\boldsymbol{\theta}^T \boldsymbol{x}_1), \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_2), \ldots, \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_N)]$.
So, for the steps of gradient descent we have

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} - \boldsymbol{X}^T (\boldsymbol{\sigma}(\boldsymbol{\theta}^T \boldsymbol{x}) - \boldsymbol{y}).$$

Solution for exercise 2

We have a dataset $Y = \{(y_i, \boldsymbol{x_i}', i = 1, \ldots, N)\}$ where $y_i \in \{0, 1\}$ is the class label for vector $\boldsymbol{x}_i' \in R^l$. The y and $\boldsymbol{x}'$ are related such as: $y = f(\boldsymbol{\theta}^T \boldsymbol{x}' + \theta_0)$ and $f(z) = 1/(1 + \exp(-az))$.

(a) We plot the function $f(z)$ fot various values of the a.



(b) The sum of error squares criterion is

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(y_i - f(\boldsymbol{\theta}^T \boldsymbol{x}_i') \right)^2$$

and the gradient with respect to $\boldsymbol{\theta}$ is

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = 2a \sum_{i=1}^{N} \left(y_i - f(\boldsymbol{\theta}^T \boldsymbol{x}_i')\right) \left(f(\boldsymbol{\theta}^T \boldsymbol{x}_i')(1 - f(\boldsymbol{\theta}^T \boldsymbol{x}_i'))\right) \boldsymbol{x_i}'$$

and we can use it for the gradient step

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{i-1}).$$

(c) We can see that for clear 1 response of the model, we need $\exp(-a\boldsymbol{\theta}^T \boldsymbol{x}_i') = 0$ which is not possible. Moreover, even if $\exp(-a\boldsymbol{\theta}^T \boldsymbol{x}_i') \to +\infty$, the model can't response a clear 0.

(d) For a given $\boldsymbol{x}$, if $f(\boldsymbol{\theta}^T \boldsymbol{x}_i')) > 0.5$, we classify it to the class $y = 1$ and for $\boldsymbol{x}$, if $f(\boldsymbol{\theta}^T \boldsymbol{x}_i')) < 0.5$, we classify it to the class $y = 0$.

(e) A way for leading the model responses very close to 1 (for class 1 vectors) or 0 (for class 0 vectors) could be to increase the parameter a to approximate step function.