Exercise for Machine Learning (SS 20)

Assignment 4: Logistic Regression

Prof. Dr. Steffen Staab, steffen.staab@ipvs.uni-stuttgart.de
Alex Baier, alex.baier@ipvs.uni-stuttgart.de
Janik Hager, janik-manel.hager@ipvs.uni-stuttgart.de
Ramin Hedeshy, ramin.hedeshy@ipvs.uni-stuttgart.de
Analytic Computing, IPVS, University of Stuttgart

Submit your solution in Ilias as either PDF for theory assignments or Jupyter notebook for practical assignments.
Mention the names of all group members and their immatriculation numbers in the file.
Submission is possible until the following Monday, 25.05.2020, at 14:00.

1 Classification with Linear Regression

Consider the following 1-dimensional input $x = [-2.0, -1.0, 0.5, 0.6, 5.0, 7.0]$ with corresponding binary class labels $y = [0, 0, 1, 0, 1, 1]$. Use (least-squares) linear regression, as shown in the lecture, to train on these samples and classify them. Your model should include an intercept term.

1. Provide the coefficients β of the linear regression (on x and y) and explain shortly how you computed them.

   Solution: Linear Regression of an indicator matrix. We define

   $$X = \begin{pmatrix} 1 & -2.0 \\ 1 & -1.0 \\ 1 & 0.5 \\ 1 & 0.6 \\ 1 & 5.0 \\ 1 & 7.0 \end{pmatrix}, \quad Y = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} \beta_0^0 & \beta_0^1 \\ \beta_1^0 & \beta_1^1 \end{pmatrix} = \begin{pmatrix} \beta^0 & \beta^1 \end{pmatrix}$$

   where the first column of B refers to the vector β^0 for the 0 class case and the second column respectively to the vector β^1 for the 1 class case.

   To compute the β matrix B, we use the corresponding formula ($B = (X^\top X)^{-1} X^\top Y$). This way, we get

   $$B = \begin{pmatrix} \frac{26307}{37565} & \frac{11258}{37565} \\ -\frac{894}{7513} & \frac{894}{7513} \end{pmatrix} \approx \begin{pmatrix} 0.7003 & 0.2997 \\ -0.1190 & 0.1190 \end{pmatrix}$$

2. Classify each of the 6 samples with your linear regression model. Explain how you map the continuous output of the linear model to a class label.

   Solution: Since we computed the β matrix B in the previous step, we can use this to predict an output with the corresponding formula ($\hat{Y} = XB$). This way, we get

$$\hat{Y} = \begin{pmatrix} \frac{35247}{37565} & \frac{2318}{37565} \\ \frac{30777}{37565} & \frac{6788}{37565} \\ \frac{24072}{37565} & \frac{13493}{37565} \\ \frac{4725}{7513} & \frac{2788}{7513} \\ \frac{3957}{37565} & \frac{33608}{37565} \\ -\frac{453}{3415} & \frac{3868}{3415} \end{pmatrix} \approx \begin{pmatrix} 0.9383 & 0.0617 \\ 0.8193 & 0.1807 \\ 0.6408 & 0.3592 \\ 0.6289 & 0.3711 \\ 0.1053 & 0.8947 \\ -0.1327 & 1.1327 \end{pmatrix}$$

   Now we can classify according to the argmax-function of the $\hat{Y}$ matrix. This gives us the following class predictions for every data point $\hat{y} = [0, 0, 0, 0, 1, 1]$. As you can see, a linear function is (of course) not able to differentiate this example correctly. An interesting side effect is that the values of the $\hat{Y}$ matrix look like probabilities (since they sum up to 1 for each row), except for the last row (as the values are either < 0 and > 1).

3. Discuss in your own words, why linear regression is not suitable for classification.

2 Log-likelihood gradient and Hessian

Consider a binary classification problem with data $D = \{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. We define

$$f(x) = \phi(x)^\top \beta \ , \quad p(x) = \sigma(f(x)) \ , \quad \sigma(z) = 1/(1 + e^{-z})$$

$$L^{\text{nll}}(\beta) = -\sum_{i=1}^{n} \Big[y_i \log p(x_i) + (1 - y_i) \log[1 - p(x_i)] \Big]$$

where $\beta \in \mathbb{R}^d$ is a vector. (Note: $p(x)$ is a short-hand for $p(y = 1|x)$.)

1. Compute the derivative $\frac{\partial}{\partial \beta} L(\beta)$. Tip: Use the fact that $\frac{\partial}{\partial z} \sigma(z) = \sigma(z)(1 - \sigma(z))$.

2. Compute the 2nd derivative $\frac{\partial^2}{\partial \beta^2} L(\beta)$.

Solution: Let $p_i \equiv p(x_i)$. We have $\frac{\partial}{\partial\beta} p_i = p_i(1-p_i)\phi(x_i)^\top$

$$
\begin{aligned}
L(\beta) &= -\sum_{i=1}^{n} \Big[y_i \log p_i + (1-y_i) \log[1-p_i] \Big] \\
\frac{\partial}{\partial\beta} L(\beta) &= -\sum_{i=1}^{n} \Big[y_i \frac{p_i(1-p_i)}{p_i} \phi(x_i)^\top + (1-y_i) \frac{-p_i(1-p_i)}{1-p_i} \phi(x_i)^\top \Big] \\
&= -\sum_{i=1}^{n} \Big[y_i(1-p_i) - (1-y_i)p_i \Big] \ \phi(x_i)^\top \\
&= \sum_{i=1}^{n} \Big[p_i - y_i \Big] \ \phi(x_i)^\top = (p-y)^\top X \\
\frac{\partial^2}{\partial\beta^2} L(\beta) &= \frac{\partial}{\partial\beta} \sum_{i=1}^{n} \phi(x_i) \Big[p_i - y_i \Big] \\
&= \sum_{i=1}^{n} \phi(x_i) p_i(1-p_i) \phi(x_i)^\top = X^\top W X \ , \quad W = \mathrm{diag}(p \circ (1-p))
\end{aligned}
$$

3 Discriminative Function in Logistic Regression

Logistic Regression defines class probabilities as proportional to the exponential of a discriminative function:

$$P(y|x) = \frac{\exp f(x,y)}{\sum_{y'} \exp f(x,y')}$$

Prove that, in the binary classification case, you can assume $f(x,0) = 0$ without loss of generality.
This results in

$$P(y=1|x) = \frac{\exp f(x,1)}{1+\exp f(x,1)} = \sigma(f(x,1)).$$

(Hint: First assume $f(x,y) = \phi(x,y)^\top\beta$, and then define a new discriminative function f' as a function of the old one, such that $f'(x,0) = 0$ and for which $P(y|x)$ maintains the same expressibility.)
Solution: Assume $f(x,y) = \phi(x,y)^\top\beta$. Define new discriminative function $f'(x,y) = f(x,y) - f(x,0)$.
Proof that new discriminative function f' still fulfills class probabilites:

$$
\begin{aligned}
P'(y|x) &= \frac{\exp f'(x,y)}{\sum_{y'} \exp f'(x,y')} \\
&= \frac{\exp(f(x,y) - f(x,0))}{\sum_{y'} \exp(f(x,y') - f(x,0))} \\
&= \frac{\exp(f(x,y)) \cdot \exp(-f(x,0))}{\sum_{y'} \exp(f(x,y')) \cdot \exp(-f(x,0))} \\
&= \frac{\exp(-f(x,0))}{\exp(-f(x,0))} \cdot \frac{\exp(f(x,y))}{\sum_{y'} \exp(f(x,y'))} \\
&= \frac{\exp f(x,y)}{\sum_{y'} \exp f(x,y')} = P(y|x)
\end{aligned}
$$