Assignment 3 SSY316

Matthew Newson, Erik Norlin

November 2023

Exercise 1

Suppose we collect data from a group of students in a Machine Learning class with variables x_1 = hours studied, x_2 = grade point average, and y = a binary output indicating whether that student received grade 5 ($y = 1$) or not ($y = 0$). We learn a logistic regression model

$$p(y = 1|x) = \frac{e^{\hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2}}{1 + e^{\hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2}}$$

with parameters $\hat{\theta}_0 = -6$, $\hat{\theta}_1 = 0.05$, and $\hat{\theta}_2 = 1$.

i)

The probability of getting a 5 using the parameters $\theta_0 = -6$, $\theta_1 = 0.05$ is

$$p(y = 1|x) = \frac{e^{\theta_0 + \theta_1 x_1 + \theta_2 x_2}}{1 + e^{\theta_0 + \theta_1 x_1 + \theta_2 x_2}} = \frac{e^{-6 + 0.05 x_1 + x_2}}{1 + e^{-6 + 0.05 x_1 + 1 x_2}}$$

Now, with $x_1 = 40$ and $x_2 = 3.5$,

$$p(y = 1|x) = \frac{e^{-6 + 0.05 \cdot 40 + 1 \cdot 3.5}}{1 + e^{-6 + 0.05 \cdot 40 + 1 \cdot 3.5}} = \frac{e^{-0.5}}{1 + e^{-0.5}}$$

$$= \frac{1}{1 + e^{0.5}} \approx 38\%.$$

ii)

Set $p(y = 1|x) = 0.5$ and $x_2 = 3.5$. This gives

$$0.5 = \frac{e^{-6 + 0.05 x_1 + 3.5}}{1 + e^{-6 + 0.05 x_1 + 3.5}} = \frac{1}{e^{2.5 - 0.05 x_1} + 1}$$

$$0.5(1 + e^{2.5 - 0.05 x_1}) = 1$$

$$e^{2.5 - 0.05 x_1} = \frac{1}{0.5} - 1 = 1$$

$$2.5 - 0.05 x_1 = \log(1) = 0$$

$$x_1 = \frac{2.5}{0.05} = 50 \,\text{hours}$$

Exercise 2

i)

The sigmoid is defined as set

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

We want to show the following

$$\frac{d\sigma(a)}{da} = \sigma(a)(1-\sigma(a))$$

This is how we show it

$$\begin{aligned}
\frac{d}{dx}\sigma(x) &= \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) && (1)\\
&= \frac{d}{dx}(1+e^{-x})^{-1} && (2)\\
&= -(1+e^{-x})^{-2}\frac{d}{dx}(e^{-x}) && (3)\\
&= -\frac{e^{-x}}{(1+e^{-x})^2} && (4)\\
&= \frac{1}{1+e^{-x}} \cdot \frac{1+e^{-x}-1}{1+e^{-x}} && (5)\\
&= \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right) && (6)\\
&= \sigma(x) \cdot (1-\sigma(x)) && (7)
\end{aligned}$$

ii)

Here we derive an expression for the gradient of the log-likelihood

$$\begin{aligned}
\mathcal{L}(\theta) = p(t_{\mathcal{D}}|x_{\mathcal{D}},\theta) &= \prod_{i=1}^{N}(t_i|x_i,\theta) = \prod_{i=1}^{N} p(\theta^T x_i)^{t_i}(1-p(\theta^T x_i))^{1-t_i}\\
\Rightarrow \mathcal{L}_{\log}(\theta) &= \sum_{1=1}^{N}\left[t_i ln(p(\theta^T x_i)) + (1-t_i) ln(1-p(\theta^T x_i))\right]
\end{aligned}$$

where $p(\theta^T x_i) = \sigma(\theta^T x_i) = \frac{1}{1+exp(-\theta^T x_i)}$, so we have that

$$\begin{aligned}
\nabla_\theta \mathcal{L}_{\log}(\theta) &= \sum_{i=1}^{N}\left[t_i \frac{p(\theta^T x_i)p(1-p(\theta^T x_i))x_i}{p(\theta^T x_i)} - (1-t_i)\frac{p(\theta^T x_i)p(1-p(\theta^T x_i))x_i}{1-p(\theta^T x_i)}\right]\\
&= \sum_{i=1}^{N}\left[t_i(1-p(\theta^T x_i))x_i - (1-t_i)p(\theta^T x_i)x_i\right]\\
&= \sum_{i=1}^{N}\left[(t_i - p(\theta^T x_i)x_i\right]\\
&= X^T(t_{\mathcal{D}} - y)
\end{aligned}$$

iii)

We want to show that the hessian $H = X^T SX$ is positive definite where $S = \text{diag}(\mu_1(1-\mu_1), ..., \mu_n(1-\mu_n)$ and $0 < \mu_i < 1$. Consider a matrix A, in theory we know that if A is diagonalizable then there exists a matrix P such that it can be written as $A = PDP^{-1}$, where D is a diagonal matrix with all the eigenvalues of A. If P is orthogonal we can write $P^{-1} = P^T$, we then have $A = PDP^T$. Since H can be written as $H = X^T SX \in R^{N\times N}$ where S is a diagonal matrix, then let $A = H$, $P = X^T$ and $D = S$. Hence, the eigenvalues of H can be found in the diagonal of S. Because of the assumption that $0 < \mu_i < 1$, $\forall i \in \{0, ..., N\}$, all eigenvalues of H are positive for all points, and hence, H is positive definite which implies that the objective function $-\mathcal{L}_{log}$ is convex. Thus, the minimum of this function is indeed the MLE.

Exercise 3

We want to create a generative binary classification model for classifying non-negative one dimensional data. This means, that the labels are binary ($y \in 0, 1$) and the samples are $x \in [0, \infty)$. We assume uniform class probabilities

$$p(y = 0) = p(y = 1) = 1/2$$

As our samples x are non-negative, we use exponential distributions (and not Gaussians) as class conditionals:

$$p(x|y = 0) = Expo(x|\lambda_0) \text{ and } p(x|y = 1) = Expo(x|\lambda_1)$$

where $\lambda_0 \neq \lambda_1$. We assume, that the parameters λ_0 and λ_1 are known and fixed

i)

The name of the posterior distribution $p(y|x)$ is the Bernoulli distribution because of the binary outcome.

ii)

Here we want to calculate what values of x classify as class 1. $x \in \mathcal{C}_1$ if $p(y = 1|x) > p(y = 0|x)$, we have that

$$\frac{p(y = 1|x)}{p(y = 0|x)} > 1, \text{ and } \frac{p(y = 0|x)}{p(y = 1|x)} > 0$$

We take the logarithm of the left inequality

$$\begin{aligned}
ln\left(\frac{p(y = 1|x)}{p(y = 0|x)}\right) &= ln\left(\frac{p(x|y = 1)p(y = 1)}{p(x|y = 0)p(y = 0)}\right) \\
&= ln\left(\frac{p(x|y = 1)}{p(x|y = 0)}\right) \\
&= ln\left(\frac{\lambda_1 e^{-\lambda_1 x}}{\lambda_0 e^{-\lambda_0 x}}\right) \\
&= ln(\lambda_1) - \lambda_1 x - ln(\lambda_0) + \lambda_0 x \\
&= x(\lambda_0 - \lambda_1) - ln(\lambda_0/\lambda_1)
\end{aligned}$$

combining the left inequality and the derived expression we get that x takes the following values if $x \in \mathcal{C}_1$

$$\begin{aligned}
x &\in \left(\frac{ln(\lambda_0/\lambda_1}{\lambda_0 - \lambda_1}, \infty\right) \text{ if } \lambda_0 > \lambda_1 \\
x &\in \left[0, \frac{ln(\lambda_0/\lambda_1}{\lambda_0 - \lambda_1}\right) \text{ otherwise}
\end{aligned}$$

Exercise 4

Here we consider a generative classification model for C classes defined by class probabilities $p(y = c) = \pi_c$ and general class-conditional densities $p(x|y = c, \theta_c)$, where $x \in \mathcal{R}^D$ is the input feature vector and $\theta = \{\theta_c\}_{c=1}^C$ are further model parameters. Suppose we are given a training set $D = \{(x^{(n)}, y^{(n)})\}_{n=1}^N$, where $y^{(n)}$ is a binary target vector of length C that uses the 1-of-C (one-hot) encoding scheme, so that it has components $y_c^{(n)} = \delta_{ck}$ if pattern n is from class $y = k$. We assume that the data points are i.i.d., and we want to show that the maximum-likelihood solution for the class probabilities π is given by $\pi_c = N_c/N$ where N_c is the number of data points assigned to class c.

To find the MLE of π_c we first define the likelihood as

$$p(D|\{\pi_c, \theta_c\}_{c=1}) = \prod_{n=1}^N \prod_{c=1}^C (p(x^{(n)}|\theta_c)\pi_c)^{y_c^{(n)}}$$

hence the log-likelihood becomes

$$\mathcal{L}_{\log} = \ln p(D|\{\pi_c, \theta_c\}_{c=1}) = \sum_{n=1}^N \sum_{c=1}^C y_c^{(n)} \log \pi_c + \text{const}$$

We maximise the log-likelihood with respect to π_c to find the MLE of π_c. To do that we lagrangian relax the constraint $\sum_{c=1}^C \pi_c = 1$

$$\sum_{n=1}^N \sum_{c=1}^C y_c^{(n)} \log \pi_c - \lambda \left(\sum_{c=1}^C \pi_c - 1 \right)$$

Then we take the derivative with respect to π_c and setting the expression to 0 which yields

$$\sum_{n=1}^N \frac{y_c^{(n)}}{\pi_c} - \lambda = 0 \Rightarrow \pi_c = \frac{1}{\lambda} \sum_{n=1}^N y_c^{(n)} = \frac{N_c}{\lambda}$$

Since we have that

$$\sum_{c=1}^C \pi_c = 1$$

we insert N_c/λ into the constraint and solve for λ

$$\sum_{c=1}^C \frac{N_c}{\lambda} = 1 \Rightarrow \lambda = N$$

Putting $\lambda = N$ into the previous expression we obtain

$$\pi_c = \frac{N_c}{N}$$

as we wanted to show.

Exercise 5

Using the same classification model as in the previous question, now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix, so that

$$p(x|y = c, \theta) = p(x|\theta_c) = N(x|\mu_c, \Sigma)$$

We want to show that the maximum likelihood estimate for the mean of the Gaussian distribution for class c is given by

$$\mu_c = \frac{1}{N_c} \sum_{\substack{n=1 \\ y^{(n)}=c}}^{N} x^{(n)}$$

which represents the mean of the observations assigned to class c

To find the MLE of μ_c we first define the log-likelihood as

$$\begin{aligned}
\mathcal{L}_{\log} &= \ln p(D|\{\pi_c, \theta_c\}_{c=1}) \\
&= \ln \left(\prod_{n=1}^{N} \prod_{c=1}^{C} (\pi_c \mathcal{N}(x^{(n)}|\mu_c, \Sigma))^{y_c^{(n)}} \right) \\
&= \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} [\ln(\pi_c) + \mathcal{N}(x^{(n)}|\mu_c, \Sigma)] \\
&= \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} \left[\ln(\pi_c) + \ln \left(\frac{1}{\sqrt{(2\pi)^D \det(\Sigma)}} e^{-1/2(x^{(n)}-\mu_c)\Sigma^{-1}(x^{(n)}-\mu_c)} \right) \right] \\
&= \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} \left[\ln(\pi_c) - \frac{1}{2}(x^{(n)} - \mu_c)\Sigma^{-1}(x^{(n)} - \mu_c) + \frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln(\det(\Sigma)) \right] \\
&= \frac{-1}{2} \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} \left[-2\ln(\pi_c) + (x^{(n)} - \mu_c)\Sigma^{-1}(x^{(n)} - \mu_c) - D\ln(2\pi) - \ln(\det(\Sigma)) \right]
\end{aligned}$$

We find the MLE by maximising the log-likelihood function with respect to μ_c by taking the derivative with respect to μ_c and setting it to 0, then solving for μ_c which yields

$$\nabla_{\mu_c} \mathcal{L}_{\log} = \frac{-1}{2} \sum_{n=1}^{N} y_c^{(n)} (2\Sigma^{-1}(\mu_c - x^{(n)})) = \sum_{n=1}^{N} y_c^{(n)} \Sigma^{-1}(x^{(n)} - \mu_c) = 0$$

$$\Rightarrow \sum_{n=1}^{N} y_c^{(n)} \Sigma^{-1} x^{(n)} = \sum_{n=1}^{N} y_c^{(n)} \Sigma^{-1} \mu_c$$

$$\mu_c = \frac{\sum_{n=1}^{N} y_c^{(n)} \Sigma^{-1} \Sigma x^{(n)}}{\sum_{n=1}^{N} y_c^{(n)}} = \frac{1}{N_c} \sum_{\substack{n=1 \\ y^{(n)}=c}}^{N} x^{(n)}$$

which we wanted to show.