Machine Learning Homework 05

Authors: Jesus Arturo Sol Navarro

1 Problem 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:

$$p(y = 0) = p(y = 1) = \frac{1}{2}.$$

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

$$p(x \mid y = 0) = \text{Exp}(x \mid \lambda_0) \quad \text{and} \quad p(x \mid y = 1) = \text{Exp}(x \mid \lambda_1),$$

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

(a) Suppose you are given an observation x. What is the name of the posterior distribution $p(y \mid x)$? You only need to provide the name of the distribution (e.g., "normal", "gamma", etc.), not estimate its parameters.

The posterior $p(y \mid x)$ is a Bernoulli distribution because $y \in \{0, 1\}$.

(b) What values of x are classified as class 1? (As usual, we assume that the classification decision is $\hat{y} = \arg\max_k p(y = k \mid x)$.)

The decision boundary is determined by solving the following inequality: $p(y = 1 \mid x) > p(y = 0 \mid x)$. This simplifies to comparing $p(x \mid y = 1)$ and $p(x \mid y = 0)$, as the priors $p(y)$ are equal:

$$\lambda_1 e^{-\lambda_1 x} > \lambda_0 e^{-\lambda_0 x}.$$

Simplifying:

$$x > \frac{\ln(\lambda_1) - \ln(\lambda_0)}{\lambda_0 - \lambda_1}, \quad \text{if } \lambda_0 > \lambda_1,$$

or:

$$x < \frac{\ln(\lambda_1) - \ln(\lambda_0)}{\lambda_0 - \lambda_1}, \quad \text{if } \lambda_0 < \lambda_1.$$

Thus, the decision boundary depends on the relationship between λ_0 and λ_1, and x is classified as class 1 if it satisfies the above inequality.

2 Problem 4

2.1 Solution

2-class classification

The posterior distribution is : $Y_i|x\ Bernoulli(\sigma(w^T x_i))$

$$p(y = 1|w.X) = \sigma(w^T x_i) p(y = 0|w.X) = 1 - \sigma(w^T x_i)$$

For the likelihood we get:

$$P(y|w.X) = \prod_{i=1}^{N} \sigma(w^T x_i)^{y!} \cdot (1 - \sigma(w^T x_i))^{1-y_i}$$

and for the negative log-likelihood:

$$-logP(y|w.X) = -\sum_{i=1}^{N} y_i \log \sigma(w^T x_i) + (1 - y_i) \log(1 - \sigma(w^T x_i))$$

For a linearly separable dataset, there exists a vector $\tilde{w}$ such that: $\tilde{w}^T x_i > 0$ if $y_i = 1$ and $\tilde{w}^T x_i < 0$ if $y_i = 0$.

$$\lim_{\lambda->\infty}(-\log P(y|\lambda w, x)) = -(\sum_{i=1}^{N} \log \lim_{\lambda->\infty} \sigma(\lambda \tilde{w}^T x_i) + \sum_{i=1}^{N} \log(1 - \lim_{\lambda->\infty} \sigma(\lambda \tilde{w}^T x_i))) = 0$$

$-\log P(y|\lambda w, x)$ is the negative sum of concave functions (log-sigmoid terms), so it is convex. A convex function has a unique minimum, our function tends towards its minimum when $\lambda- > \infty$, so $-\log P(y|\lambda w, x)$ has an infinite minimizer: $||w||- > \infty$.
To prevent $\|\mathbf{w}\| \to \infty$, we add a convex regularization term $\lambda\|\mathbf{w}\|^2$ to the objective function. This penalizes large $\|\mathbf{w}\|$, ensuring a finite solution.

3 Problem 5

3.1 Solution

$$\begin{aligned}
\text{Softmax:}\quad p(y=1|\mathbf{x}) &= \frac{\exp(\mathbf{w}_1^\top \mathbf{x})}{\exp(\mathbf{w}_1^\top \mathbf{x}) + \exp(\mathbf{w}_0^\top \mathbf{x})} \\
&= \frac{1}{1 + \frac{\exp(\mathbf{w}_0^\top \mathbf{x})}{\exp(\mathbf{w}_1^\top \mathbf{x})}} \\
&= \frac{1}{1 + \exp(\mathbf{w}_0^\top \mathbf{x} - \mathbf{w}_1^\top \mathbf{x})} \\
&= \frac{1}{1 + \exp(-(\mathbf{w}_1 - \mathbf{w}_0)^\top \mathbf{x})} \\
&= \sigma(\tilde{\mathbf{w}}^\top \mathbf{x}) \quad \text{where } \tilde{\mathbf{w}} = \mathbf{w}_1 - \mathbf{w}_0
\end{aligned}$$

4 Problem 6

4.1 Solution

Show that the derivative of the sigmoid function $\sigma(a) = (1 + e^{-a})^{-1}$ can be written as

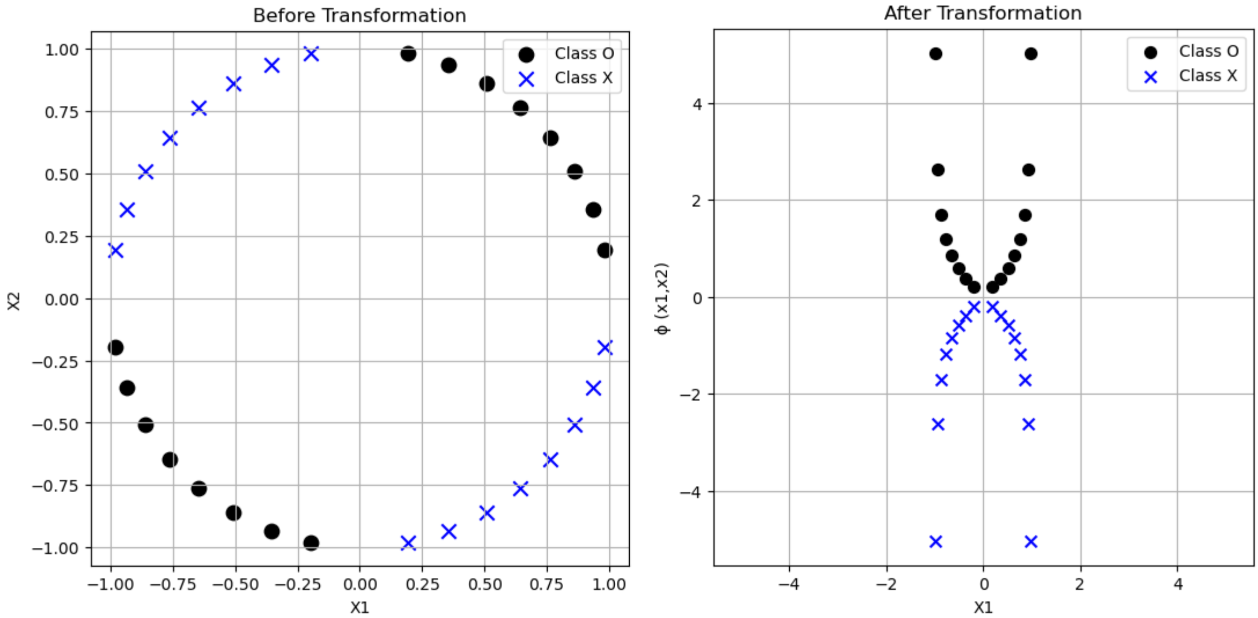
$$\frac{\partial \sigma(a)}{\partial a} = \sigma(a)\,(1 - \sigma(a))\,.$$

Differentiating w.r.t a

$$\frac{\partial \sigma(a)}{\partial a} = \frac{e^{-a}}{(1+e^{-a})^2} = \frac{1 - 1 + e^{-a}}{(1+e^{-a})^2} = \frac{1}{1+e^{-a}}\left(1 - \frac{1}{1+e^{-a}}\right) = \sigma(a)\,(1 - \sigma(a))$$

5 Problem 7

Give a basis function $\phi(x_1, x_2)$ that makes the data in the example below linearly separable



The basis function $\phi(x_1, x_2) = (x_1, \frac{x_1}{x_2})$ creates a new feature space where the relationship between x_1 and x_2 is transformed. This transformation helps align the data into a pattern that can be separated by a straight line in the new space.

6 Problem 8

6.1 Solution

$$P(y=1 \mid X) + P(y=0 \mid X) = 1 \quad \text{and} \quad P(y=1 \mid X) + P(y=0 \mid X) = 1$$

$$\frac{P(y=1 \mid X)}{P(y=0 \mid X)} = 1$$

$$\log \frac{P(y=1 \mid X)}{P(y=0 \mid X)} = \log \left(\frac{P(X \mid y=1)P(y=1)}{P(X)} \cdot \frac{P(X)}{P(X \mid y=0)P(y=0)} \right)$$

$$= \log\left(P(X \mid y=1) \cdot \pi_1\right) - \log\left(P(X \mid y=0) \cdot \pi_0\right)$$

$$= \log N(X \mid \mu_1, S_1) - \log N(X \mid \mu_0, S_0) + \log \frac{\pi_1}{\pi_0}$$

$$= -\frac{1}{2}\left(\log(2\pi|S_1|) + (x-\mu_1)^T S_1^{-1}(x-\mu_1)\right) + \frac{1}{2}\left(\log(2\pi|S_0|) + (x-\mu_0)^T S_0^{-1}(x-\mu_0)\right) + \log\frac{\pi_1}{\pi_0}$$

$$= -\frac{1}{2}x^T S_1^{-1} x + x^T S_1^{-1}\mu_1 - \frac{1}{2}\mu_1^T S_1^{-1}\mu_1 + \frac{1}{2}x^T S_0^{-1} x - x^T S_0^{-1}\mu_0 + \frac{1}{2}\mu_0^T S_0^{-1}\mu_0 + \frac{1}{2}\log|S_0| - \frac{1}{2}\log|S_1| + \log\frac{\pi_1}{\pi_0}$$

$$= x^T \frac{1}{2}[S_0^{-1} - S_1^{-1}]x + x^T[S_1^{-1}\mu_1 - S_0^{-1}\mu_0] - \frac{1}{2}\mu_1^T S_1^{-1}\mu_1 + \frac{1}{2}\mu_0^T S_0^{-1}\mu_0 + \log\frac{\pi_1}{\pi_0} + \frac{1}{2}\log\frac{|S_0|}{|S_1|}$$

Let:

$$A = \frac{1}{2}[S_0^{-1} - S_1^{-1}], \quad b = S_1^{-1}\mu_1 - S_0^{-1}\mu_0, \quad c = -\frac{1}{2}\mu_1^T S_1^{-1}\mu_1 + \frac{1}{2}\mu_0^T S_0^{-1}\mu_0 + \log\frac{\pi_1}{\pi_0} + \frac{1}{2}\log\frac{|S_0|}{|S_1|}$$

The decision boundary is:

$$\Gamma = \{x \mid x^T A x + b^T x + c = 0\}$$