Universidad Nacional de Colombia
Departarmento de Matemáticas

Matemáticas para el Aprendizaje de Máquina

Juan Antonio Rodríguez

2024

Problems Section 5

1. Problem 5.1 Show that the logistic sigmoid function $sig[z]$ maps $z = -\infty$ to 0, $z = 0$ to 0.5 and $z = \infty$ to 1 where:

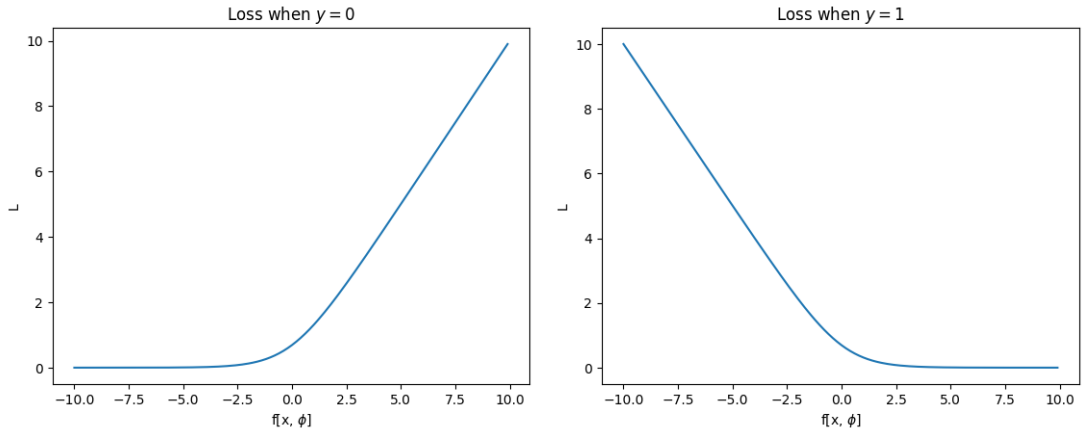
$$sig[z] = \frac{1}{1 + exp[-z]}$$

- $z = -\infty$.
  $\lim_{z\to-\infty} \frac{1}{1+exp[-z]} = \frac{1}{1+exp[\infty]} = \frac{1}{\infty} = 0.$
  So, $\lim_{z\to-\infty} \frac{1}{1+exp[-z]} = 0$
- $z = 0$
  $\lim_{z\to0} \frac{1}{1+exp[-z]} = \frac{1}{1+exp[0]} = \frac{1}{2} = \frac{1}{2}.$
  So, $\lim_{z\to0} \frac{1}{1+exp[-z]} = \frac{1}{2}$
- $z = \infty$
  $\lim_{z\to\infty} \frac{1}{1+exp[-z]} = \frac{1}{1+exp[-\infty]} = \frac{1}{1+0} = 1.$
  So, $\lim_{z\to\infty} \frac{1}{1+exp[-z]} = 1$

2. Problem 5.2 The loss L for binary classification for a single training pair x, y is:

$$L = -(1-y)log[1 - sig[f[x,\phi]]] - ylog[sig[f[x,\phi]]],$$

where $sig[\cdot]$ is defined in equation 5.32. Plot this loss as a function of the transformed network output $sig[f[x,\phi]] \in [0,1]$ (i) when the training label $y = 0$ and (ii) when $y = 1$.



3. Problem 5.3 Suppose we want to build a model that predicts the direction y in radians of the prevailing wind based on local measurements of barometric pressure x. A suitable distribution over circular domains is the von Mises distribution (figure 5.13):

$$Pr(y|\mu,\kappa) = \frac{exp[\kappa cos[y-\mu]]}{2\pi Bessel_0[\kappa]}$$

where μ is a measure of the mean direction and κ is a measure of the concentration (i.e., the inverse of the variance). The term $Bessel_0[\kappa]$ is a modified Bessel function of order 0. Use the recipe from section 5.2 to develop a loss function for learning the parameter μ of a model $f[x,\phi]$ to predict the most likely wind direction. Your solution should treat the concentration κ as constant. How would you perform inference?

(a) Probability Distribution: Von Mises Distribution.

$$Pr(y|\mu,\kappa) = \frac{exp[\kappa cos[y-\mu]]}{2\pi Bessel_0[\kappa]}$$

Parameters: μ and κ.

(b) Model will predict μ.

(c) Negative Log Likelihood:

$$\begin{aligned}\hat{\phi} &= \underset{\phi}{\text{argmin}}\left[-\sum_{i=1}^{I} log\left[Pr(y_i|f(x_i,\phi))\right]\right] \\ &= \underset{\phi}{\text{argmin}}\left[-\sum_{i=1}^{I} log\left[\frac{exp[\kappa cos[y_i - f[x_i,\phi]]]}{2\pi Bessel_0[\kappa]}\right]\right]\end{aligned}$$

$$L[\phi] = -\sum_{i=1}^{I} log\left[\frac{exp[\kappa cos[y - f[x,\phi]]]}{2\pi Bessel_0[\kappa]}\right]$$

When performing inference, the model can return the mean predicted direction μ and if needed the value of the concentration κ which corresponds to the inverse of the variance. Returning both values is the way the model has to not only predict the direction but also to give a value of certainty or margin of error.

4. Problem 5.4. Sometimes, the outputs y for input x are multimodal (figure 5.14a); there is more than one valid prediction for a given input. Here, we might use a weighted sum of normal components as the distribution over the output. This is known as a mixture of Gaussians model. For example, a mixture of two Gaussians has parameters $\theta = \lambda, \mu_1, \sigma^2, \mu_2, \sigma^2$:

$$Pr(y|\lambda, \mu_1, \sigma^2, \mu_2, \sigma^2) = \frac{\lambda}{\sqrt{2\pi\sigma^2}} exp[\frac{-(y-\mu_1)^2}{2\sigma^2}] + \frac{1-\lambda}{2\pi\sigma^2} exp[\frac{-(y-\mu_2)^2}{2\sigma^2}]$$

where $\lambda \in [0, 1]$ controls the relative weight of the two components, which have means μ_1, μ_2 and variances σ^2, σ^2, respectively. This model can represent a distribution with two peaks (figure 5.14b) or a distribution with one peak but a more complex shape (figure 5.14c). Use the recipe from section 5.2 to construct a loss function for training a model $f[x, \phi]$ that takes input x, has parameters ϕ, and predicts a mixture of two Gaussians. The loss should be based on I training data pairs $\{x_i, y_i\}$. What problems do you foresee when performing inference?

(a) Probability Distribution: Mixture of two Gaussians.

$$Pr(y|\lambda, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \frac{\lambda}{\sqrt{2\pi\sigma_1^2}} exp[\frac{-(y-\mu_1)^2}{2\sigma_1^2}] + \frac{1-\lambda}{2\pi\sigma_2^2} exp[\frac{-(y-\mu_2)^2}{2\sigma_2^2}]$$

Parameters: μ_1, μ_2, σ_1, σ_2 and λ.

(b) Model will predict all paramters.

- $\mu_1 = f_1[x, \phi]$
- $\mu_2 = f_2[x, \phi]$
- $\sigma_1 = f_3[x, \phi]$
- $\sigma_2 = f_4[x, \phi]$
- $\lambda = f_5[x, \phi]$

(c) Negative Log Likelihood:

$$\begin{aligned}\hat{\phi} &= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log\left[Pr(y_i|f(x_i, \phi))\right] \right] \\ &= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log\left[\frac{f_5[x_i, \phi]}{\sqrt{2\pi f_3[x_i, \phi]^2}} exp[\frac{-(y_i - f_1[x_i, \phi])^2}{2f_3[x_i, \phi]^2}] + \frac{1-f_5[x_i, \phi]}{2\pi f_4[x_i, \phi]^2} exp[\frac{-(y_i - f_2[x_i, \phi])^2}{2f_4[x_i, \phi]^2}] \right] \right]\end{aligned}$$

$$L[\phi] = -\sum_{i=1}^{I} log\left[\frac{f_5[x, \phi]}{\sqrt{2\pi f_3[x, \phi]^2}} exp[\frac{-(y - f_1[x, \phi])^2}{2f_3[x, \phi]^2}] + \frac{1-f_5[x, \phi]}{2\pi f_4[x, \phi]^2} exp[\frac{-(y - f_2[x, \phi])^2}{2f_4[x, \phi]^2}] \right]$$

Problems one may foresee when performing inference:

When performing inference and the mixture of two Gaussians result in a graph with two peaks, if both means have the same value, inference may result in a matter of picking one or the other without reasoning.

Also, when returing not only the mean value in some x, but also the variance (or a value of certainty), the model output would get complicated if the generated probability graph is one with only one peak but a complicated shape.

5. Problem 5.5 Consider extending the model from problem 5.3 to predict the wind direction using a mixture of two von Mises distributions. Write an expression for the likelihood $Pr(y|\theta)$ for this model. How many outputs will the network need to produce?

   (a) Probability Distribution: Mixture of two Von Mises.

   $$Pr(y|\lambda, \mu_1, \kappa_1, \mu_2, \kappa_2) = \frac{exp[\kappa cos[y-\mu]]}{2\pi Bessel_0[\kappa]}$$

   Parameters: μ_1, μ_2, κ_1, κ_2 and λ.

   (b) Model will predict all paramters.

   - $\mu_1 = f_1[x, \phi]$
   - $\mu_2 = f_2[x, \phi]$
   - $\kappa_1 = f_3[x, \phi]$
   - $\kappa_2 = f_4[x, \phi]$
   - $\lambda = f_5[x, \phi]$

   (c) Negative Log Likelihood:

   $$\begin{aligned}\hat{\phi} &= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log\left[Pr(y_i|f(x_i,\phi))\right] \right] \\ &= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log\left[(f_5[x,\phi]) \frac{exp[f_3[x,\phi]cos[y-f_1[x,\phi]]}{2\pi Bessel_0[f_3[x,\phi]]} \right.\right. \\ &\left.\left. +(1-f_5[x,\phi]) \frac{exp[f_4[x,\phi]cos[y-f_2[x,\phi]]]}{2\pi Bessel_0[f_4[x,\phi]]} \right] \right]\end{aligned}$$

   $$\begin{aligned}L[\phi] = -\sum_{i=1}^{I} log\left[(f_5[x,\phi]) \frac{exp[f_3[x,\phi]cos[y-f_1[x,\phi]]}{2\pi Bessel_0[f_3[x,\phi]]} \right. \\ \left. +(1-f_5[x,\phi]) \frac{exp[f_4[x,\phi]cos[y-f_2[x,\phi]]]}{2\pi Bessel_0[f_4[x,\phi]]} \right]\end{aligned}$$

6. Problem 5.6 Consider building a model to predict the number of pedestrians $y \in \{0, 1, 2, \dots\}$ that will pass a given point in the city in the next minute, based on data x that contains information about the time of day, the longitude and latitude, and the type of neighborhood. A suitable distribution for modeling counts is the Poisson distribution (figure 5.15). This

has a single parameter $\lambda > 0$ called the rate that represents the mean of the distribution. The distribution has probability density function:

$$Pr(y = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Design a loss function for this model assuming we have access to I training pairs x_i, y_i.

(a) Probability Distribution: Poisson.

$$Pr(y = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Parameters: λ.

(b) Model will predict λ.

(c) Negative Log Likelihood:

$$\begin{aligned}\hat{\phi} &= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log \left[Pr(y_i = k|f(x_i, \phi)) \right] \right] \\ &= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log \left[\frac{(f[x, \phi])^k e^{-f[x,\phi]}}{k!} \right] \right]\end{aligned}$$

$$L[\phi] = -\sum_{i=1}^{I} log \left[\frac{(f[x, \phi])^k e^{-f[x,\phi]}}{k!} \right]$$

7. Problem 5.7 Consider a multivariate regression problem where we predict ten outputs, so $y \in \mathbb{R}^{10}$, and model each with an independent normal distribution where the means μ_d are predicted by the network, and variances σ^2 are constant. Write an expression for the likelihood $Pr(y|f[x, \phi])$. Show that minimizing the negative log-likelihood of this model is still equivalent to minimizing a sum of squared terms if we don't estimate the variance σ^2.

(a) Probability Distribution: 10 Independent Normal.

$$Pr(y|f[x, \phi]) = \prod_d Pr(y_d|f_d[x, \phi])$$

Where,

$$Pr_d(y_d|f_d[x, \phi]) = \frac{1}{\sqrt{2\pi\sigma_d^2}} exp[\frac{-(y_d - \mu_d)^2}{2\sigma_d^2}$$

Parameters: μ_d.

(b) Model will predict all μ_d.

(c) Negative Log Likelihood:

$$\begin{aligned}
\hat{\phi} &= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log\left[Pr(y_i|f(x_i,\phi))\right] \right] \\
&= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log\left[\prod_d Pr(y_{i_d}|f_d[x,\phi]) \right] \right] \\
&= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log\left[\prod_d \frac{1}{\sqrt{2\pi\sigma_d^2}} exp[\frac{-(y_d - f_d[x,\phi])^2}{2\sigma_d^2} \right] \right] \\
&= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} \sum_d log\left[\frac{1}{\sqrt{2\pi\sigma_d^2}} exp[\frac{-(y_d - f_d[x,\phi])^2}{2\sigma_d^2} \right] \right]
\end{aligned}$$

$$L[\phi] = -\sum_{i=1}^{I} \sum_d log\left[\frac{1}{\sqrt{2\pi\sigma_d^2}} exp[\frac{-(y_d - f_d[x,\phi])^2}{2\sigma_d^2} \right]$$

8. Problem 5.8 Construct a loss function for making multivariate predictions $y \in \mathbb{R}^{D_i}$ based on independent normal distributions with different variances σ^2 for each dimension. Assume a heteroscedastic model so that both the means μ_d and variances σ^2 vary as a function of the data.

(a) Probability Distribution: 10 Independent Normal.

$$Pr(y|f[x,\phi]) = \prod_d Pr(y_d|f_d[x,\phi])$$

Where,

$$Pr_d(y_d|f_d[x,\phi]) = \frac{1}{\sqrt{2\pi\sigma_d^2}} exp[\frac{-(y_d - \mu_d)^2}{2\sigma_d^2}$$

Parameters: μ_d.

(b) Model will predict all μ_d and σ_d, where $f_{2d}[x,\phi] = \mu_d$ and $f_{2d+1}[x,\phi] = \sigma_d$

(c) Negative Log Likelihood:

$$\hat{\phi} = \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log\left[Pr(y_i|f(x_i,\phi))\right] \right]$$
$$= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log\left[\prod_{d} Pr(y_{i_d}|f_d[x,\phi]) \right] \right]$$
$$= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log\left[\prod_{d} \frac{1}{\sqrt{2\pi f_{2d+1}[x,\phi]^2}} exp[\frac{-(y_d - f_{2d}[x,\phi])^2}{2f_{2d+1}[x,\phi]^2} \right] \right]$$
$$= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} \sum_{d} log\left[\frac{1}{\sqrt{2\pi f_{2d+1}[x,\phi]^2}} exp[\frac{-(y_d - f_{2d}[x,\phi])^2}{2f_{2d+1}[x,\phi]^2} \right] \right]$$

$$L[\phi] = -\sum_{i=1}^{I} \sum_{d} log\left[\frac{1}{\sqrt{2\pi f_{2d+1}[x,\phi]^2}} exp[\frac{-(y_d - f_{2d}[x,\phi])^2}{2f_{2d+1}[x,\phi]^2} \right]$$

9. Problem 5.9 Consider a multivariate regression problem in which we predict the height of a person in meters and their weight in kilos from data x. Here, the units take quite different ranges. What problems do you see this causing? Propose two solutions to these problems.

   The main consecuence of the outputs having quite different ranges is that it would be harder to fit the model due to both outputs using the same parameters in the network.

   Two possible solutions to this problem would be:

   (a) Doing pre-processing in the dataset where the height would be converted from meters to centimeters and the weight would be also converted, from kilograms pounds. Human anatomy and this units yield similar ranges. It would be easier to fit a model using this units. When doing inference, it just requiere an addicional conversion back to meters for the height and kilograms for the weight.

   (b) Using more hidden units, that being, more hidden units per layer or more layers. Making this network more robust and sensitive to this changes in outout range. However, this solution would require a lot more computing power and memory, due to having a lot more parameters that need to be stored and fit.

10. Problem 5.10 Extend the model from problem 5.3 to predict both the wind direction and the wind speed and define the associated loss function.

    (a) Probability Distribution: Von Mises and Normal, independent from each other.

$$Pr_1(y|\mu_1,\kappa) = \frac{exp[\kappa cos[y-\mu_1]]}{2\pi Bessel_0[\kappa]}$$
$$Pr_2(y|\mu_2,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp[\frac{-(y_d-\mu_2)^2}{2\sigma^2}$$

Parameters: μ_1, μ_2, σ and κ.

(b) Model will predict all parameters:

- $f_1[x, \phi] = \mu_1$
- $f_2[x, \phi] = \mu_2$
- $f_3[x, \phi] = \sigma$
- $f_4[x, \phi] = \kappa$

(c) Negative Log Likelihood:

$$\begin{aligned}
\hat{\phi} &= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log \left[\prod_{j=1}^{2} Pr_j(y_i | f(x_i, \phi)) \right] \right] \\
&= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log \left[\left[\frac{exp[f_4[x,\phi] cos[y - f_1[x,\phi]]]}{2\pi Bessel_0[f_4[x,\phi]]} \right] \left[\frac{1}{\sqrt{2\pi f_3[x,\phi]^2}} exp[\frac{-(y_d - f_2[x,\phi])^2}{2 f_3[x,\phi]^2} \right] \right] \right] \\
&= \underset{\phi}{\text{argmin}} \left[-\sum_{i=1}^{I} log \left[\frac{exp[f_4[x,\phi] cos[y - f_1[x,\phi]]]}{2\pi Bessel_0[f_4[x,\phi]]} \right] + log \left[\frac{1}{\sqrt{2\pi f_3[x,\phi]^2}} exp[\frac{-(y_d - f_2[x,\phi])^2}{2 f_3[x,\phi]^2} \right] \right]
\end{aligned}$$

$$L[\phi] = -\sum_{i=1}^{I} log \left[\frac{exp[f_4[x,\phi] cos[y - f_1[x,\phi]]]}{2\pi Bessel_0[f_4[x,\phi]]} \right] + log \left[\frac{1}{\sqrt{2\pi f_3[x,\phi]^2}} exp[\frac{-(y_d - f_2[x,\phi])^2}{2 f_3[x,\phi]^2} \right]$$