5 Loss Functions

Problem 5.1 Show that the logistic sigmoid function $\text{sig}[z]$ becomes 0 as $z \to -\infty$, is 0.5 when $z = 0$, and becomes 1 when $z \to \infty$, where:

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

- $z \to -\infty$: $\exp[-z] \to \infty$, so $\text{sig}[z] \to 0$.
- $z = 0$: $\text{sig}[z] = 1/(1+1) = 0.5$.
- $z \to \infty$: $\exp[-z] \to 0$, so $\text{sig}[z] \to 1$.

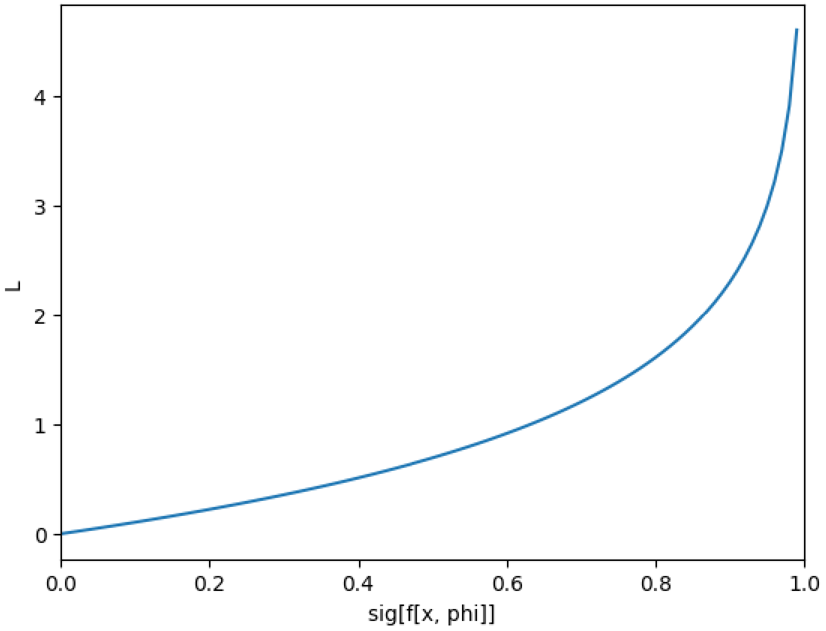
Problem 5.2 The loss L for binary classification for a single training pair $\{\mathbf{x}, y\}$ is:

$$L = -(1-y)\log\left[1 - \text{sig}[f[\mathbf{x}, \boldsymbol{\phi}]]\right] - y\log\left[\text{sig}[f[\mathbf{x}, \boldsymbol{\phi}]]\right]$$

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

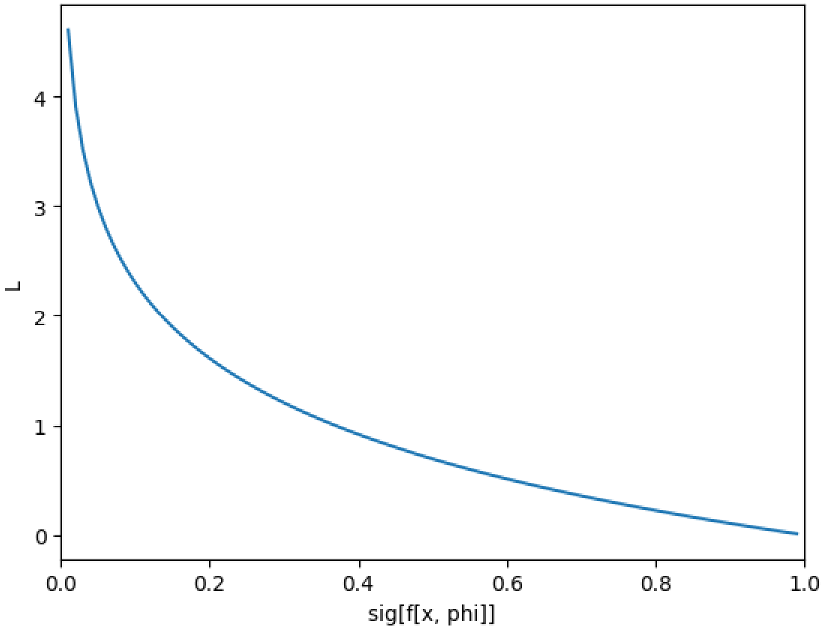
(i) $y = 0$:

$$L = -\log[1 - \text{sig}[f[\mathbf{x}, \boldsymbol{\phi}]]]$$



(ii) $y = 1$:

$$L = -\log[\text{sig}[f[\mathbf{x}, \boldsymbol{\phi}]]$$



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 $\mathbf{x}$. A suitable distribution over circular domains is the von Mises distribution (figure 5.13):

$$Pr(y|\mu,\kappa) = \frac{\exp[\kappa\cos[y-\mu]]}{2\pi\cdot\text{Bessel}_0[\kappa]}$$

where μ is a measure of the mean direction and κ is a measure of concentration (i.e. the inverse of the variance). The term $\text{Bessel}_0[\kappa]$ is a modified Bessel function of the first kind of order 0. Use the recipe from section 5.2 to develop a loss function for learning the parameter μ of a model $f[\mathbf{x},\boldsymbol{\phi}]$ to predict the most likely wind direction. Your solution should treat the concentration κ as constant.. How would you perform inference?

Setup that the model $f[\mathbf{x},\boldsymbol{\phi}]$ predicts μ, where we can minimize the negative log likelihood loss function over training dataset pairs $\{\mathbf{x}_i, y_i\}$:

$$\begin{aligned}\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\arg\min}[-\sum_i \log Pr(y_i|f[\mathbf{x}_i,\boldsymbol{\phi}])] \\ &= \underset{\boldsymbol{\phi}}{\arg\min}[-\sum_i [\kappa\cos[y_i - f[\mathbf{x}_i,\boldsymbol{\phi}]] - C]] \\ &= \underset{\boldsymbol{\phi}}{\arg\min} -\sum_i \cos[y_i - f[\mathbf{x}_i,\boldsymbol{\phi}]]\end{aligned}$$

Thus we can define the loss function as

$$L = \sum_i -\cos[y_i - f[\mathbf{x}_i,\boldsymbol{\phi}]]$$

Problem 5.4 Sometimes, the outputs y for the input $\mathbf{x}$ is 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 $\boldsymbol{\theta} = \{\lambda, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2\}$:

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

where $\lambda \in [0,1]$ controls the relative weight of the two components, which have means μ_1, μ_2 and variances σ_1^2, σ_2^2, 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 $\mathbf{f}[x,\boldsymbol{\phi}]$ that takes input x, has parameters $\boldsymbol{\phi}$, 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?

The loss function is the negative log-likelihood based on the conditional probability produced by the model prediction:

$$
\begin{aligned}
\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\arg\min}[-\sum_i \log Pr(y_i|\mathbf{f}[x_i, \boldsymbol{\phi}])] \\
&= \underset{\boldsymbol{\phi}}{\arg\min}[-\sum_i \log[\frac{f_1[x_i, \boldsymbol{\phi}]}{\sqrt{2\pi f_3[x_i, \boldsymbol{\phi}]}} \exp[\frac{-(y_i - f_2[x_i, \boldsymbol{\phi}]_1)^2}{2 f_3[x_i, \boldsymbol{\phi}]}] \\
&+ \frac{1 - f_1[x_i, \boldsymbol{\phi}]}{\sqrt{2\pi f_5[x_i, \boldsymbol{\phi}]}} \exp[\frac{-(y_i - f_4[x_i, \boldsymbol{\phi}]_2)^2}{2 f_5[x_i, \boldsymbol{\phi}]}]]]]
\end{aligned}
$$

Even if the model is based on multiple Gaussian components, naïve inference based on maximal probability only outputs predictions on one of the two peaks.

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 $Pr(y|\boldsymbol{\theta})$ for this model. How many outputs will the network need to produce?

A mix of two von Mises distributions with parameters $\boldsymbol{\theta} = \{\mu_1, \kappa_1, \mu_2, \kappa_2\}$ has a distribution of:

$$
\begin{gathered}
Pr(y|\mu_1, \kappa_1, \mu_2, \kappa_2) \\
= \lambda \frac{\exp[\kappa_1 \cos[y - \mu_1]]}{2\pi \cdot \mathrm{Bessel}_0[\kappa_1]} + (1 - \lambda) \frac{\exp[\kappa_2 \cos[y - \mu_2]]}{2\pi \cdot \mathrm{Bessel}_0[\kappa_2]}
\end{gathered}
$$

and the network would need to produce 4 outputs as learned parameters.

Problem 5.6 Consider building a model to predict the number of pedestrians $y \in \{0, 1, 2, \cdots\}$ that will pass a given point in the city in the next minute, based on data $\mathbf{x}$ 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 $\{\mathbf{x}_i, y_i\}$.

Setup the model $f[\mathbf{x}, \boldsymbol{\phi}]$ which predicts the model parameter λ:

$$\begin{aligned}\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\arg\min} - \sum_i \log Pr(y = y_i) \\ &= \underset{\boldsymbol{\phi}}{\arg\min} - \sum_i \log \frac{f[\mathbf{x}, \boldsymbol{\phi}]^{y_i} e^{-f[\mathbf{x}, \boldsymbol{\phi}]}}{y_i!} \\ &= \underset{\boldsymbol{\phi}}{\arg\min} \sum_i [f[\mathbf{x}, \boldsymbol{\phi}] - y_i \log f[\mathbf{x}, \boldsymbol{\phi}]]\end{aligned}$$

Therefore, the loss function can be defined by

$$L = f[\mathbf{x}, \boldsymbol{\phi}] - y_i \log f[\mathbf{x}, \boldsymbol{\phi}]$$

Problem 5.7 Consider a multivariate regression problem where we predict ten outputs, so $\mathbf{y} \in \mathbb{R}^{10}$, 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 $Pr(\mathbf{y}|\mathbf{f}[\mathbf{x}, \boldsymbol{\phi}])$. Show that minimizing the negative log-likelihood of this model is still equivalent to minimizing a sum of square terms if we don't estimate the variance σ^2.

Given that each outputs are independent, the likelihood is the product of the individual likelihoods in each of the output. Thus, the negative log-likelihood is again the sum of individual log-likelihoods.

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\arg\min} \sum_i \sum_d \frac{(y_{i,d} - f_d[\mathbf{x}_i, \boldsymbol{\phi}])^2}{2\sigma^2}$$

If we regard the variance $\sigma_d^2 = \sigma^2$ constant for all dimensions, then the expression turns back to the sum of square errors.

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

In this case, we setup the output predicts $2d$ results, where $f_{d,1}[\mathbf{x}, \boldsymbol{\phi}]$ predicts the mean of the d-th dimension μ_d, and $f_{d,2}[\mathbf{x}, \boldsymbol{\phi}]$ predicts the variance σ_d^2.

Minimizing the negative log-likelihood gives:

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\arg\min} \sum_i - \log[\prod_d \frac{1}{\sqrt{2\pi f_{d,2}[\mathbf{x}, \boldsymbol{\phi}]}} \exp[\frac{(y_i - f_{d,1}[\mathbf{x}, \boldsymbol{\phi}])^2}{2f_{d,2}[\mathbf{x}, \boldsymbol{\phi}]}]]$$

which we can rewrite with

$$\underset{\boldsymbol{\phi}}{\arg\min} \sum_i \sum_d [\log \sqrt{2\pi f_{d,2}[\mathbf{x}, \boldsymbol{\phi}]} - \frac{(y_i - f_{d,1}[\mathbf{x}, \boldsymbol{\phi}])^2}{2f_{d,2}[\mathbf{x}, \boldsymbol{\phi}]}]]$$

Thus, the loss function can be defined as

$$L = \sum_i \sum_d [\log \sqrt{2\pi f_{d,2}[\mathbf{x}, \boldsymbol{\phi}]} - \frac{(y_i - f_{d,1}[\mathbf{x}, \boldsymbol{\phi}])^2}{2f_{d,2}[\mathbf{x}, \boldsymbol{\phi}]}]]$$

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 $\mathbf{x}$. Here, the units take quite different ranges. What problems do you see this causing? Propose two solutions to these problems.

Suppose we regard two dimensions independent and use the sum of squared errors approach. Note that sum of squared errors assume that the variance for each output dimension is the same: this is where the problem arises. Since the scale of two dimensions are different, a dimension (probably weight) will have the dominating error term. To solve this problem, we could normalize the input scale of both dimensions using the statistics from input or prior knowledge. Other way is to add the coefficient to the error term, which basically does the same thing. The variance of both dimensions can be predicted separately so that the model can regard the scale differences.

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.

Assuming that wind direction and speed are independent, we can combine the loss function by simply adding the loss function for von Mises distribution (for wind direction) and loss function for Gaussian (for wind speed).

This model predicts four parameters: $\boldsymbol{\theta} = \{\mu, \kappa, \lambda, \sigma^2\}$.

$$L = -\sum_i \log \frac{\exp[f_2[\mathbf{x}, \boldsymbol{\phi}] \cos[y_{i,1} - f_1[\mathbf{x}, \boldsymbol{\phi}]]]}{2\pi \cdot \text{Bessel}_0[f_2[\mathbf{x}, \boldsymbol{\phi}]]} + \log \frac{1}{\sqrt{2\pi f_4[\mathbf{x}, \boldsymbol{\phi}]}} \exp[\frac{(y_{i,2} - f_3[\mathbf{x}, \boldsymbol{\phi}])^2}{2f_4[\mathbf{x}, \boldsymbol{\phi}]}]$$