Chapter 8: Measuring performance

Problem 8.1

Will the multiclass cross-entropy training loss in figure 8.2 ever reach zero? Explain your reasoning.

Under practical conditions, it is unlikely that the multiclass cross-entropy training loss in figure 8.2 will ever reach zero. As the softmax function is commonly used in the output layer for multiclass classification problems to convert the raw class scores (logits) into probabilities. Due to the exponential function involved in softmax, the predicted probabilities for incorrect classes are unlikely to be exactly zero, even for a very confident model.

Problem 8.2

What values should we choose for the three weights and biases in the first layer of the model in figure 8.4a so that the hidden unit's responses are as depicted in figures 8.4b-d?

For 8.4b, $\theta_{10} = 0, \text{and} \theta_{11} = 1$, for 8.4c, $\theta_{20} = -0.3, \text{and} \theta_{21} = 1$, and for 8.4d, $\theta_{30} = -0.7$ and $\theta_{31} = 1$.

Problem 8.3

Given a training set of I input and output pairs $\{x_i, y_i\}$, show how the parameters $\{\beta, \omega_1, \omega_2, \omega_3, \}$ for the model in figure 8.4a using the least squares loss function can be found in closed form.

The key is that the first part of the network is now deterministic; we can compute the activations at the three hidden units for any input. Denoting these by h1, h2 and h3, we now have a linear regression problem:

$$y_i = \beta + \omega_1 h_{1i} + \omega_2 h_{2i} + \omega_3 h_{3i}$$

where i indexes the training data.

Problem 8.4

Consider the curve in figure 8.10b at the point where we train a model with a hidden layer of size 200, which would have 50,410 parameters. What do you predict will happen to the training and test performance if we increase the number of training examples from 10,000

to 50,410?

Increasing the number of training examples from 10,000 to 50,410 (increasing training dataset size) would traditionally reduce the variance and therefore improve the generalization performance of the model. This is because the model would have more data to learn from, which would help it to generalize better to unseen data. Therefore, we would expect the test performance to improve.

Problem 8.5

Consider the case where the model capacity exceeds the number of training data points, and the model is flexible enough to reduce the training loss to zero. What are the implications of this for fitting a heteroscedastic model? Propose a method to resolve any problems that you identify

Heteroscedasticity refers to the condition where the variance of the errors differs across the range of values of an independent variable. The most immediate implication would be that the model overfits to the training data and therefore generalizes poorly to unseen data. To resolve this, we could introduce a regularization term to the loss function to prevent the model from overfitting to the training data or use cross-validation techniques to tune model hyperparameters and select the model complexity that generalizes best to unseen data.

Problem 8.6

Show that two random points drawn from a 1000-dimensional standard Gaussian distribution are orthogonal relative to the origin with high probability.

For two vectors to be orthogonal, their dot product must be zero. The dot product of two random samples X and Y from a 1000-dimensional standard Gaussian distribution is given by:

$$x \cdot y = \sum_{i=1}^{1000} X_i Y_i$$

Since X_i and Y_i are independent and identically distributed (i.i.d) with mean 0 and variance of 1, the expected value of $X_i Y_i$ is also zero and $\text{Var}(X_i Y_i) = 1$. By the Central Limit Theorem (CLT), as the dimensionality grows, the distribution of the sum (normalized by its standard deviation) will approach a standard normal distribution.

The variance of the sum is the sum of the variances (since the variables are independent):

$$\text{Var}(x \cdot y) = \sum_{i=1}^{1000} \text{Var}(X_i Y_i) = 1000$$

The standard deviation of $X \cdot Y$ is therefore $\sqrt{1000}$. Therefore, as the dimensionality increases, the normalized dot product $(X \cdot Y)/\sqrt{1000}$ will, with high probability, fall within a few standard deviations from 0.

Problem 8.7

The volume of a hypersphere with radius r in$D$ dimensions is:

$$\text{Vol}[r] = \frac{r^D \pi^{D/2}}{\Gamma(D/2+1)}$$

where $\Gamma\bullet$ is the gamma function. Show using Stirling's formula that the volume of a hypersphere of diameter one ($r = 0.5$) becomes zero as the dimension increases.

Stirling's formula approximates the gamma function for large x as:

$$\Gamma(x) \approx \sqrt{2\pi} x^{x-0.5} e^{-x}$$

Substituting this into the volume formula:

$$\begin{aligned}\text{Vol}[0.5] &\approx \frac{(0.5)^D \pi^{D/2}}{\sqrt{2\pi}(D/2+1)^{D/2} e^{-D/2}} \\ &\approx \pi^{\frac{D}{2}-\frac{1}{2}} (0.5\sqrt{2})^D (D+2)^{-\frac{D}{2}-\frac{1}{2}} e^{\frac{D}{2}+1}\end{aligned}$$

The key factor here is the exponential decay caused by the $(0.5\sqrt{2})^D$ term. As D increases, this term diminishes much faster than the other components can grow, leading to the overall volume approaching zero.

Problem 8.8

Consider a hypersphere of radius r = 1. Find an expression for the proportion of the total volume that lies in the outermost 1% of the distance from the center (i.e., in the outermost shell of thickness 0.01). Show that this becomes one as the dimension increases.

Volume of hypersphere of radius $r = 1$ is given by:

$$\text{Vol}[1] = \frac{\pi^{D/2}}{\Gamma(D/2+1)}$$

Volume of hypersphere with radius $r = 0.99$ is given by:

$$\text{Vol}[0.99] = \frac{(0.99)^D \pi^{D/2}}{\Gamma(D/2+1)}$$

The proportion p in the last one percent is hence:

$$p = \frac{\text{Vol}[1] - \text{Vol}[0.99]}{\text{Vol}[1]} = 1 - 0.99^D$$

Which tends to one as $D \to \infty$.

Problem 8.9

Figure 8.13c shows the distribution of distances of samples of a standard normal distribution as the dimension increases. Empirically verify this finding by sampling from the standard normal distributions in 25, 100, and 500 dimensions and plotting a histogram of the distances from the center. What closed-form probability distribution describes these distances?

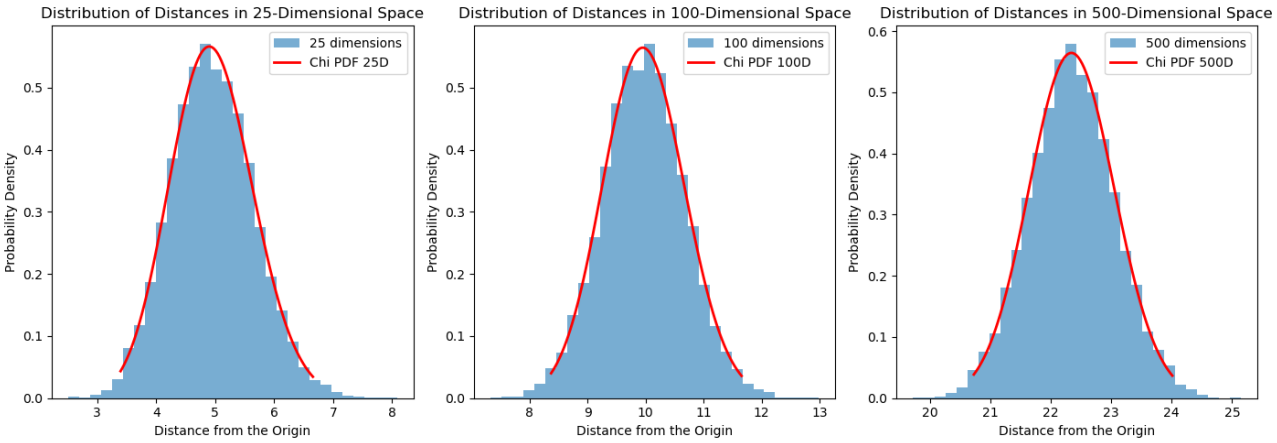

Figure 1: Distribution of distances of samples of a standard normal distribution as the dimension increases.

The distances appear to follow a normal distribution more closely as the dimensionality increases, this is a reflection of the distances concentrating around their mean due to the dimensions' contribution to the sum of squares.