1 Logistic Regression

Assume that we have n i.i.d. data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, where each y_i is a binary label in $\{0, 1\}$. We model the posterior probability as a Bernoulli distribution and the probability for each class is the sigmoid function, i.e., $p(y|\mathbf{x}; \mathbf{w}) = q^y(1-q)^{1-y}$, where $q = s(\mathbf{w}^\top \mathbf{x})$ and $s(z) = \frac{1}{1+\exp\{-z\}}$ is the sigmoid function.

(a) **Show that the derivative of the sigmoid function is:** $s'(z) = s(z)(1 - s(z))$

(b) **Write out the likelihood and log likelihood functions, along with the MLE objective.** Comment on whether it is possible to find a closed form maximum likelihood estimate of $\mathbf{w}$, and describe an alternate approach.

(c) **Write the stochastic gradient descent update step with learning rate η, where the gradient step is calculated on a single data point $(\mathbf{x}_i, y_i)$.**

2 Gaussian Classification: LDA and QDA

Gaussian discriminant analysis (GDA) is a generative classification method, which involves modeling the posterior probability by approximating the underlying class-conditional data distribution $p_\theta(\mathbf{x} \mid y)$ and class priors $p_\theta(y)$. We call this "generative" because we model the *generating* distribution of the data.

The fundamental assumption that GDA makes is that the class-conditional data distribution is Gaussian, and the priors over classes form a Bernoulli distribution (or a multinomial distribution with > 2 classes):

$$p_\theta(\mathbf{x} \mid y = C_i) \sim \mathcal{N}(\mu_i, \mathbf{\Sigma}_i)$$

$$p_\theta(y) \sim \text{Bernoulli}(\pi)$$

The three steps for performing GDA are as follows:

1. Find the parameters of the Gaussian class-conditional data distribution and the prior probabilities using MLE on the (labeled) dataset.

2. Combine the two distributions to produce a quantity proportional to the posterior:

$$p_\theta(y \mid \mathbf{x}) \propto p_\theta(\mathbf{x} \mid y) p_\theta(y)$$

3. Construct a classifier to determine the class of an input point based on the class with the maximum posterior probability:

$$\text{Classifier}_{\text{GDA}}(\mathbf{x}) = \underset{C_1,\ldots,C_k}{\arg\max}\ p_\theta(Y = C_i \mid \mathbf{x})$$

We will focus the binary classification case in this worksheet. However, one nice property of GDA is that it scales up to multi-class classification very easily.

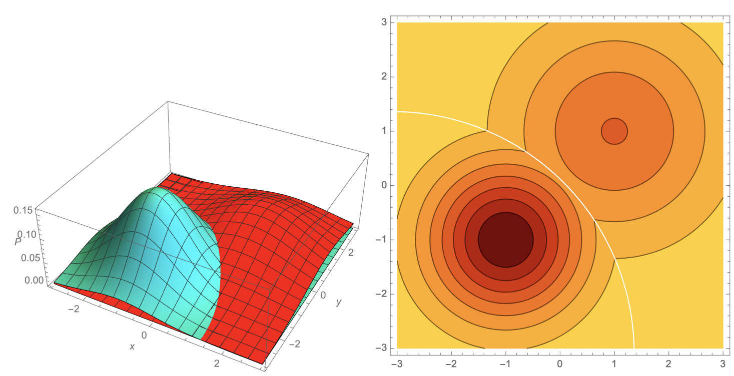

Figure 1: Figure taken from Professor Shewchuck's notes

(a) (Step 1) **Given a set of points $\mathcal{D}_i = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_i}\}$ labeled as the class C_i, what are the MLE estimates for the mean μ_i and covariance matrix $\mathbf{\Sigma}_i$ for this class?**

(b) (Step 1) **Express the MLE estimates of the priors $\pi_0 = p_\theta(Y = 0), \pi_1 = p_\theta(Y = 1)$ in terms of the number of data points in each class n_0, n_1.**

(c) (Steps 2-3) **Write an equation describing the decision boundary where the posterior probabilities are equal.** Leave it as a quadratic form, no need to simplify fully.

$$p_{\theta_0=(\mu_0,\Sigma_0)}(Y = 0 \mid \mathbf{x}) = p_{\theta_1=(\mu_1,\Sigma_1)}(Y = 1 \mid \mathbf{x})$$

As a reminder, the PDF for a d-dimensional multivariate Gaussian is:

$$f(\mathbf{x} \in \mathbb{R}^d) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right\}.$$

(d) The algorithm developed in the previous parts is known as Quadratic Discriminant Analysis (QDA). In lecture, we primarily focused on a simpler variant called Linear Discriminant Analysis (LDA), where each class is assumed to have the *same* covariance matrix $\mathbf{\Sigma}$ rather than modeling many separate covariance matrices.

Assuming a shared covariance matrix Σ, write the decision boundary as a linear function in the form $\mathbf{w}^\top \mathbf{x} + b$. Identify $\mathbf{w}$ and b.

$$p_{\theta_0=(\mu_0,\Sigma)}(Y = 0 \mid \mathbf{x}) = p_{\theta_1=(\mu_1,\Sigma)}(Y = 1 \mid \mathbf{x})$$

(e) **When plugging in the fitted distributions for LDA, what is the posterior probability $p_\theta(Y = 1 \mid \mathbf{x})$?**

Hint: Recall the formula for the sigmoid function $s(z) = \frac{1}{1+\exp\{-z\}}$.

Hint: You should be able to reuse a lot of work from the previous part!

(f) The Bayes optimal classifier (given a symmetric loss function) is one that classifies any point $\mathbf{x}$ as the class with maximum posterior probability $p(Y = C_i \mid \mathbf{x})$:

$$\text{Classifier}_{\text{Bayes}}(\mathbf{x}) = \underset{C_1,\ldots,C_k}{\arg\max}\ p(Y = C_i \mid \mathbf{x})$$

Why is the decision boundary found by LDA or QDA not generally the Bayes optimal decision boundary?

Hint: What assumptions did we make in GDA?

