CS 189: Introduction to Machine Learning - Discussion 5

1. Logistic Posterior with different variances

In class, we've discussed the decision boundary obtained from two Gaussian class conditionals with different variances. Now we will derive the decision boundary for this case, described as

$$X|Y = i \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad \text{where } i \in \{0, 1\}$$
$$Y \sim \text{Bernoulli}(\pi)$$

Show that the posterior distribution of the class label given X looks *like* a logistic, however with a quadratic argument in X. Assuming 0-1 loss, what will the decision boundary look like (i.e., describe what the posterior probability plot looks like)? What name have we given this method?

Solution: This is the derivation of the decision boundary for Quadratic Discriminant Analysis.

We are solving for $\mathbb{P}(Y = 1|x)$. By Bayes Rule, we have

$$\begin{aligned}
\mathbb{P}(Y = 1|x) &= \frac{\mathbb{P}(x|Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(x|Y = 1)\mathbb{P}(Y = 1) + \mathbb{P}(x|Y = 0)\mathbb{P}(Y = 0)} \\
&= \frac{1}{1 + \frac{\mathbb{P}(Y=0)\mathbb{P}(x|Y=0)}{\mathbb{P}(Y=1)\mathbb{P}(x|Y=1)}} \\
&= \frac{1}{1 + \frac{\sigma_1}{\sigma_0}\frac{1-\pi}{\pi}\exp\left(\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_0)^2}{2\sigma_0^2}\right)}
\end{aligned}$$

Looking at the bottom right equation, we have

$$\frac{\sigma_1}{\sigma_0}\frac{1-\pi}{\pi}\exp\left(\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_0)^2}{2\sigma_0^2}\right)$$

$$= \exp\left[\left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_0^2}\right)x^2 + \left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right)x + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_0^2}{\sigma_0^2} + \ln\left(\frac{\sigma_1}{\sigma_0}\frac{1-\pi}{\pi}\right)\right)\right]$$

Now we see that we have a logistic function $\frac{1}{1+\exp(-h(x))}$, where $h(x) = ax^2 + bx + c$, for appropriate values of a, b, c, is a quadratic function. Note that the special case examined in class of $\sigma_1 = \sigma_0$ gives a linear function in x.

Since we are assuming 0-1 loss, we use the optimal classifier $f^*(x) = 1$ when $\mathbb{P}(Y = 1|x) > \mathbb{P}(Y = 0|x)$. Thus, the decision boundary can be found when $\mathbb{P}(Y = 1|x) = \mathbb{P}(Y = 0|x) = \frac{1}{2}$. This happens when $h(x) = 0$. Solving for the roots of $h(x)$ results in 2 values where this equality holds. One can convince themself that in the plot of posterior probability graph, the horizontal (x) axis will be split into three regions: we classify the two outer regions as one class, and the middle one as another class. The choice of which class to classify in the outer regions depends on the values of σ_1 and σ_2.

2. Gradient Descent for Linear Regression

In class, we've seen the normal equation solution to linear regression. Here, we will work through an alternate method for solving linear regression problems. We will use the same loss function, the squared loss, that we have been using.

$$L = \sum_{i=1}^{n}(y_i - w^T x_i)^2$$

a. Write the generic update equation for batch gradient descent (untied to linear regression).

b. Write the generic update equation for stochastic gradient descent.

c. Write **both** a batch gradient descent and a stochastic gradient descent update step equation for solving linear regression.

d. Why might we use gradient descent optimization (numerical) rather than the normal equations (analytical)?
When might we use each solution method (analytical vs. numerical)?

Solution:

a. $w^{t+1} = w^t - \alpha * \sum_{i=1}^{n} \frac{dL_i(w)}{dw}$, where $L_i(w)$ represents the ith term in the loss function.

b. $w^{t+1} = w^t - \alpha * \frac{dL_i(w)}{dw}$, where $L_i(w)$ represents the ith term in the loss function.

c. $\frac{dL_i(w)}{dw} = \frac{d}{dw}(y_i - w^T x_i)^2$
$= \frac{d}{dw} y_i^2 - 2y_i w^T x_i + (w^T x_i)^2$
$= -2y_i * x_i + 2x_i * x_i^T * w$

Stochastic Update: $w^{t+1} = w^t - \alpha(-2y_i * x_i + 2x_i * x_i^T * w)$

Batch Update: $w^{t+1} = w^t - \alpha \sum_{i=1}^{n} -2y_i * x_i + 2x_i * x_i^T * w$

d. To solve the analytical solution, we need to invert the matrix $X^T X$, which is an $O(n^3)$ operation–this can grow expensive quite quickly. As a result, we might get to the solution quicker using gradient descent. Since the squared loss function is convex, we can obtain the correct (global minimum) solution.

3. Probabilistic Formulation of Linear Regression

So far, we've considered linear regression from a linear algebra/geometric perspective. In this question, we will develop a probabilistic interpretation of linear regression, based on our work with maximum likelihood estimation. Rather than starting with a squared loss function, we will instead start with the assumption that our data comes from a line, $w^T x$, but is modified by additive Gaussian noise.

$$\text{Noise: } \epsilon = \mathcal{N}(0, \sigma^2)$$
$$y = w^T x + \epsilon$$

Take a moment to convince yourself that this equivalent to the following distribution:

$$P(y|\mathbf{x}, \sigma^2) \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2)$$

Now, show that finding the maximum likelihood estimate for this distribution is equivalent to minimizing the squared loss.

Solution: $P(y_1, ..., y_n|\mathbf{x}, \sigma^2) \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2) = \Pi_{i=1}^n \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2)$
$= \Pi_{i=1}^n \frac{1}{\sqrt{2*\pi}*\sigma} \exp\left(-\frac{1}{2} * \frac{(y_i - w^T x_i)^2}{\sigma^2}\right)$
Taking the log: $-n \log(\sqrt{2 * \pi} * \sigma) - \sum_{i=1}^n (\frac{1}{2} * \frac{(y_i - w^T x_i)^2}{\sigma^2})$

Since sigma is constant for each sample, we can also remove sigma from the sum:
$-n \log(\sqrt{2 * \pi} * \sigma) - \frac{n}{\sigma^2} \sum_{i=1}^n (\frac{1}{2} * (y_i - w^T x_i)^2)$

This leaves our expression for the log-likelihood in a familiar form! The first term will have no effect on the solution, since it has no dependence on w. We can also ignore the constant factor in front of the sum. When optimizing over w, this term will also have no effect.

4. Visualization of Decision Boundaries + Overfitting

We will now take a look at the decision boundaries associated with the various types of regression we've seen in class, and consider the cases in which they overfit.
These visualizations will be projected, and will also be posted to Piazza later.