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?

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)?

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.

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.