Discussion 6 **Alex Francis**
CS 189 Email: afrancis@berkeley.edu
02/21/2017 OH: Wednesday 10:00am - 12:00pm (283E Soda)

1 Lecture on Regression

Recall from lecture that a regression problem, or "fitting a curve to data consists" of choosing three things.

1. A regression function. That is, a model by which we assume the data is generated. Some choices:
   (i) Linear: $h(x; w, \alpha) = w \cdot x + \alpha$
   (ii) Polynomial: As before, linear regression on polynomial feature vectors.
   (iii) Logistic: $h(x; w, \alpha) = s(w \cdot x + \alpha), s(\gamma) = \dfrac{1}{1 + e^{-\gamma}}$
2. A loss function to optimize: let z be the prediction $h(x)$; y is the true value.
   (A) Squared Error: $L(z, y) = (z - y)^2$
   (B) Absolute Error: $L(z, y) = |z - y|$
   (C) Logistic Loss: $L(z, y) = -y \ln z - (1 - y) \ln(1 - z)$
3. A way of combining loss functions over all data points, called a *cost function*.
   (a) Mean loss: $J(h) = \dfrac{1}{n} \sum_{i=1}^{n} L(h(X_i), y_i)$
   (b) Maximum loss: $J(h) = \max_i L(h(X_i), y_i)$
   (c) Weighted sum: $J(h) = \sum_{i=1}^{n} \omega_i L(h(X_i), y_i)$
   (d) ℓ_2 penalized/regularized: $J(h) = \dfrac{1}{n} \sum_{i=1}^{n} L(h(X_i), y_i) + \lambda \|w\|^2$
   (e) ℓ_1 penalized/regularized: $J(h) = \dfrac{1}{n} \sum_{i=1}^{n} L(h(X_i), y_i) + \lambda \|w\|_1$

There's something kind of beautiful and simple about this formulation. This menu tells you just about everything you need to know. Examine data; generate hypothesis about the model for the data; fit that model by minimizing cost!

The next few weeks spend time on the reasons for using different combinations of these three categories, as well as the optimization methods that are used to minimize cost in practice. We start with the simplest of examples.

2 Lecture on Least Squares and Statistical Properties

In the linear least-squares ((i), (A), (a)) formulation of the regression problem. Assume, for now, that we're examining *simple regression* (one explanatory variable X). Statisticians assume that the data is drawn from a model of the form,

$$Y_i = X_i w + \alpha + \epsilon_i$$

This is the form of a linear model, with an error term, or noise random variable. Statisticians assume, for a variety of historical reasons, that measurement error can be captured using the normal distribution as an error curve (the Laplace distribution may also be reasonable in it's place, the the normal distribution is preferred in statistics for technical reasons). Concretely, the full set assumptions are,

- Zero mean. $\mathbb{E}(\epsilon_i) = 0$.
- Constant variance. $\text{Var}(\epsilon_i) = \sigma^2$.
- Normally distributed.
- Independence. $\forall i \neq j, \text{Cov}(\epsilon_i, \epsilon_j) = 0$.

Therefore, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (Illustrate). So, $Y_i = X_i w + \alpha + \epsilon_i \sim \mathcal{N}(X_i w + \alpha, \sigma^2)$. (Do you see why? Possible derivation). Then, observe that the log-likelihood function for the data is,

$$\begin{aligned}
\ell(X_i w + \alpha, \sigma^2) &= \sum_{i=1}^{n} \log p(y_i \mid \mu, \sigma) \\
&= \sum_{i=1}^{n} \log \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y_i - x_i w + \alpha}{\sigma} \right)^2 \right) \right) \\
&= -n \log(\sigma\sqrt{2\pi}) - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - x_i w + \alpha}{\sigma} \right)^2 \\
&= -n \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i w + \alpha)^2
\end{aligned}$$

When I maximize log-likelihood with respect to w, the first term, falls out, and the multiplying constant of the sum is irrelevant. So, I am attempting to minimize a sum of square errors with respect to w. More general formulations exist for any "hypothesis" choice in the above menu. There are also analogous model formulations and results in higher dimension. The takeaway here is that maximizing the log-likelihood is the same as minimizing the cost function we've chosen for this problem — a statistical justification for the method of least squares linear regression! Least squares linear regression has a variety of other wonderful qualities, which are not limited to these two:

- The estimators returned by least-squares are *unbiased*. That is $\mathbb{E}(\hat{w}) = w$ and $\mathbb{E}(\hat{\alpha}) = \alpha$.
- Of all linear unbiased estimators, the least-squares estimators have the lowest variance (are the most precise). See bias-variance decomposition in an upcoming lecture.

3 Lecture on Logistic Regression

Recall the methods of Gaussian Discriminant Analysis. First, we would model the distribution of the data as Gaussian for each class. Upon encountering a new point, we would return the class that maximized the posterior probability. Instead of fitting a probability distribution first, let's consider an alternative route of modeling posterior probabilities *directly*. That is, let,

$$\begin{aligned}
P(Y = 1 \mid X = x) &= f(w^\top x + \alpha) \\
P(Y = 0 \mid X = x) &= 1 - f(w^\top x + \alpha)
\end{aligned}$$

The function $f(x) \in [0,1] \forall x$, so the logistic, or sigmoid function seems like a natural choice. Then,

$$P(Y \mid X = x) = s(w^\top x + \alpha)^y (1 - s(w^\top x + \alpha))^{1-y}$$

The question arises: what cost function seems like a good idea? Blindly using the same trick as we used for linear-least-squares regression, the log-likelihood function for n data points is,

$$\begin{aligned}
\ell(w, \alpha) &= \sum_{i=1}^{n} \log p(y_i \mid w, \alpha) \\
&= \sum_{i=1}^{n} \log \left(s(w^\top x + \alpha)^y (1 - s(w^\top x + \alpha))^{1-y} \right) \\
&= \sum_{i=1}^{n} y_i \log(s(w^\top x + \alpha)) + (1 - y_i) \log(1 - s(w^\top x + \alpha))
\end{aligned}$$

This is the logistic loss function.

4 Gradient Descent for Linear Regression

Mostly definitions here. See the staff solutions.

5 Newton's Method for Root Finding

1. Write down the iterative update equation of Newton's method for finding a root $x : f(x) = 0$ for a real-values function f.

   Recall the basic principle of Newton's method from 61A, Math1B, or high school calculus: approximate a function with linear Taylor Approximation. Then, compute the root of that linear function using primitive methods. This gives the derivation,

   $$f(x) \approx f(x_i) + \frac{f'(x_i)}{1!}(x - x_i)$$

   Then set,

   $$\begin{aligned}
   f(x) &= 0 \\
   f(x_i) + \frac{f'(x_i)}{1!}(x - x_i) &= 0 \\
   f'(x_i)x &= f'(x_i)x_i - f(x_i) \\
   x &= \frac{f'(x_i)x_i - f(x_i)}{f'(x_i)} \\
   &= x_i - \frac{f(x_i)}{f'(x_i)}
   \end{aligned}$$

   Therefore, at an iteration i, the update equation is,

   $$x_{i+1} \leftarrow x_i - \frac{f(x_i)}{f'(x_i)}$$

2. Prove that if $f(x)$ is quadratic ($f(x) = ax^2 + bx + c$), then it only takes one iteration of Newton's Method to find the minimum/maximum.

   Recall what Newton's method does. It fits a quadratic to the curve at a particular value using a quadratic Taylor series approximation. Then, it compute the minimum of the quadratic. From this simple description, a simple derivation is yielded. At some point (during the ith iteration of the algorithm) x_i,

   $$f(x) \approx f(x_i) + \frac{f'(x_i)}{1!}(x - x_i) + \frac{f''(x_i)}{2!}(x - x_i)^2$$

   It can be shown that the minimum of a polynomial is $-b/2a$, so the minimum of this polynomial is,

   $$-\frac{f'(x_i) - f''(x_i)x_i}{f''(x_i)} = x_i - \frac{f'(x_i)}{f''(x_i)}$$

   The proof proceeds from the update equation.

   Proof. For a polynomial $f(x) = ax^2 + bx + c$, the update equation above gives,

   $$x_{i+1} \leftarrow x_i - \frac{f'(x_i)}{f''(x_i)} = x_i - \frac{2ax_i + b}{2a} = x_i - x_i - b/2a = -b/2a$$

   □

6 Logistic Regression

Recall that the loss function for logistic regression is

$$L = \sum_{i=1}^{m} (y_i \log s(x_i^\top w) + (1 - y_i) \log(1 - s(x_i^\top w)))$$

where s is the sigmoid function.

1. Write down the batch gradient descent update for logistic regression.

   Start the problem by writing down the objective. We write the general form of the batch gradient update as,

   $$w^{t+1} \leftarrow w^t - \eta \sum_{i=1}^{n} \nabla_w L$$

   Then, computing the important term,

   $$\begin{aligned}\frac{\partial L}{\partial w} &= \sum_{i=1}^{n} \frac{\partial}{\partial w} \left(y_i \log s(x_i^\top w) + (1 - y_i) \log(1 - s(x_i^\top w)) \right) \\ &= \sum_{i=1}^{n} \left(\frac{y_i}{s(x_i^\top w)} + \frac{1 - y_i}{1 - s(x_i^\top w)} \right) s'(x_i^\top w)\end{aligned}$$

We have that $s'(\gamma) = s(\gamma)(1 - s(\gamma))$ (confirm this!). Therefore,

$$\begin{aligned}
\frac{\partial L}{\partial w} &= \sum_{i=1}^{n} \left(\frac{y_i}{s(x_i^\top w)} + \frac{1 - y_i}{1 - s(x_i^\top w)} \right) s'(x_i^\top w) \\
&= \sum_{i=1}^{n} \left(\frac{y_i}{s(x_i^\top w)} + \frac{1 - y_i}{1 - s(x_i^\top w)} \right) s(x_i^\top w)(1 - s(x_i^\top w))x_i \\
&= \sum_{i=1}^{n} (y_i - s(x_i^\top w))x_i
\end{aligned}$$

Then, the gradient descent update is,

$$w^{t+1} \leftarrow w^t - \eta \sum_{i=1}^{n} (y_i - s(x_i^\top w))x_i$$

2. Given the length of the lecture today, and the length of the worksheet, we didn't make it here!