Chris Piech
CS109

Handout #34
May 11th, 2016

Parameter Estimation

Parameters

Consider the following probability distributions:

$\text{Ber}(p)$	$\theta = p$
$\text{Poi}(\lambda)$	$\theta = \lambda$
$\text{Uni}(a,b)$	$\theta = (a,b)$
$\text{N}(\mu,\sigma^2)$	$\theta = (\mu,\sigma^2)$
$Y = mX + b$	$\theta = (m,b)$

Given a model, the parameters yield the actual distribution. We usually refer to all parameters of a distribution or a machine learning model as θ. In the real world you don't know the "true" parameters, but you get to observe data. In we can use that data to estimate the model parameters we can start to understand how the system we are modelling works – and we can begin to make predictions.

Method of Moments

Recall that we had defined a sample mean and sample variance for estimating parameters. Are they the only way to estimate parameters? No! Another way to estimate parameters is to start with the axiom that we want to chose model parameters by equating the "true" moments of the distribution to the sample moments: $m_i \approx \hat{m}_i$.

First recall what that the n-th moment of a distribution for variable X is:

$$m_n = E[X^n]$$

Consider IID random variables $X_1, X_2, \ldots X_n$. We can compute sample moments for the variables as such:

$$\hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{m}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\ldots$$

$$\hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

What does this look like for an estimate of variance?

$$\begin{aligned}\text{Var}(X) &= E[x^2] - (E[X])^2 \\ &\approx \hat{m}_2 - (\hat{m}_1)^2\end{aligned}$$

How does it relate to sample mean / variance

The "unbiased" sample mean estimator is the exact same as the method of moments estimator for mean. They diverge in their estimate of variance. Recall the sample variance was:

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

With a little algebra we can express

$$S^2 = \frac{n}{n-1}(\hat{m}_2 - (\hat{m}_1)^2)$$

Where $(\hat{m}_2 - (\hat{m}_1)^2)$ is the method of moments estimate of variance. That difference becomes especially notable when $n = 1$. In that case the sample variance is undefined, and the method of moments estimate is 1. For large values of n, where large is greater than around 30, the two measures converge.

Rich Passengers on the Titanic

Lets do a simple example that uses the dataset of Titanic passengers. Let S be a random indicator variable which is 1 if a passenger survived the Titanic sinking and X is a random variable that represents the fare paid in British Pounds to ride the Titanic. Calculate $P(S = \text{ true } | X \geq 100)$.

For notation reasons lets define Y to be the event that $X \geq 30$. If we could collect IID random samples: $S_1|Y, S_2|Y, \ldots, S_n|Y$ then we could assume $S|Y \sim \text{Ber}(p)$ and we could use either the sample mean or the method of moments to estimate p.

But where can we get IID samples? From the dataset. Consider each datapoint that is consistent with the event that $X \geq 100$. The values of S for those datapoints are samples of $S|Y$.

$$p = E[S_i|Y] \approx \hat{m}_1 = \bar{S}|Y = \frac{1}{n}\sum_{i=1}^{n} S_i|Y = \hat{p}$$
$$= \frac{39}{53} = 0.74$$

Self Driving Car Redux

An autonomous car has an instrument for determining its direction. The instrument reports a direction $D = T + X$ where $X \sim N(\mu = 0, \sigma^2 = 1)$ is Gaussian noise and T is the true direction of the autonomous car. Before checking the instrument, the car believes that all directions in the range 50 to 60 degrees are equally likely. What is the probability density function for the true direction given that the instrument reports 57 degrees (D = 57)? Use a constant K in your function.

This problem required a few steps. First use Bayes theorem to rewrite the probability density function. Then recognize that two of the three terms are constant. Finally plug in the density functions for the instrument reading given the true direction.

$$\begin{aligned}
f(T = t|D = 57) &= \frac{f(D = 57|T)f(T)}{f(D = 57)} && \text{Bayes Theorem} \\
&= Kf(D = 57|T = t) && \text{The prior is uniform. The denominator is constant.} \\
&= K \cdot \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(57-t)^2}{2}} \right) && \text{plug in the normal pdf} \\
&= K \cdot e^{-\frac{(57-t)^2}{2}}
\end{aligned}$$

There are a few ways to figure out the form of $f(D = 57|T = t)$. The one used above requires recognizing that adding a constant to a normal random variable produces a new normal with a shifted mean. The other approach is to plug in the equation for the D variable and solve for the noise term X. The noise term is defined.