Sampling and Estimation

Brian J. Mann

Estimation

Objectives

Expected Value

Recall that the *expected value* of a discrete random variable is the weighted sum:

$$E[X] = P(X = x_1) * x_1 + P(X = x_2) * x_2 + \cdots + P(X = x_n) * x_n$$

For a continuous random variable with density function f:

$$E[X] = \int x f(x)\, dx$$

Variance (1/2)

The *variance* of a random variable X is the expected value of the square difference from the mean:

$$Var(X) = E[(X - E[X])^2]$$
$$= E[X^2 - 2XE[X] + E[X]^2]$$
$$= E[X^2] - E[X]^2$$

Variance (2/2)

For a discrete random variable:

$$Var(X) = \sum_i P(X = x_i) * (x_i - E[X])^2$$

For a continuous random variable with density f:

$$Var(X) = \int (x - E[X])^2 f(x) dx$$

Inference

Parametric

- Assumes the data is drawn from a class of distributions determined by numeric parameters
- For example $Norm(\mu, \sigma)$, $Poisson(\lambda)$, or $Binom(n, p)$
- Determine which parameters are the best fit for the data

Non-Parametric

- Make no assumption about the family of distribution the data is drawn from
- More flexible
- Less interpretable, often hard to compute anything about the inferred distribution

Maximum Likelihood Estimation (MLE) (Parametric)

Assume each data point is drawn independently from the same distribution with density $f(x|\theta)$. Since the draws are independent the joint density function is

$$f(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta) * f(x_2|\theta) * \cdots * f(x_n|\theta)$$

If we have a formula for f in terms of the parameters θ, we can find the values of theta which maximizes the *likelihood*

$$\mathcal{L}(\theta|x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta) = \prod f(x_i|\theta)$$

or equivalently the *log-likelihood*

$$\log \mathcal{L}(\theta|x_1, x_2, \ldots, x_n) = \sum \log f(x_i|\theta)$$

$$\hat{\theta} = argmax_\theta \log f(x_1, x_2, \ldots, x_n|\theta)$$

Example - MLE

Suppose we flip a coin N times and get H heads. We want an estimate for how biased the coin is. Each flip is a Bernoulli trial with parameter p. The joint distribution is $Binom(N, p)$, so we need to find p which minimizes

$$\log p^H (1-p)^{N-H}$$

Maximum A Posteriori (MAP) (Parametric)

Generalization of MLE where we assume some prior distribution on the parameters θ

$$\mathcal{L}(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int_{\Theta} f(x|t)g(t)dt}$$

To find the optimal θ we find

$$\hat{\theta} = argmax_{\theta} \frac{f(x|\theta)g(\theta)}{\int_{\Theta} f(x|t)g(t)dt} = argmax_{\theta} f(x|\theta)g(\theta)$$

To get MLE, assume a uniform prior on θ so that the function g disappears from the *argmax* above

Method of Moments (MOM) (Parametric)

Older method, generally MLE is preferred. But good to know anyway.

- A *moment* of a distribution is $E[X], E[X^2], E[X^3], \ldots$
- $E[X]$ is the first moment, $E[X^2]$ is the second moment, etc...
- Use the moments to derive as many equations as parameters, and then solve

Example - MOM (1/2)

Suppose we flip a coin N times again, and get H heads. Let's use MOM this time to estimate p, the probability of flipping a head. Since the number of heads of N flips is modeled by a Binomial distribution we can compute the first moment

$$E[X] = Np$$

Since we have a single unknown, we stop at the first moment. We compute the sample first moment $\bar{x} = H$ and set this equal to theoretical first moment

$$H = Np$$

So we estimate

$$\hat{p} = H/N$$

Example - MOM (2/2)

Suppose we have data sampled from a symmetric uniform distribution with unknown bounds $X \sim Unif(-b, b)$. The first moment is

$$E[X] = 0$$

so that doesn't help. The second moment is

$$E[X^2] = Var(X) + E[X]^2 = Var(X) = b^2/3$$

Computing the sample variance s^2

$$s^2 = b^2/3$$

so that

$$\hat{b} = \sqrt{3s^2}$$

Kernel Density Estimation (KDE) (Non-Parametric)

A *kernel* is another word for a density function of a distribution with mean 0.

Kernel Density Estimation estimates a distribution empirically given data by summing kernels centered at each point. The density function of the kernel density estimate is:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_i K\left(\frac{x - x_i}{h}\right)$$

The parameter h is called the *bandwidth*, and it's analogous to the width of bins in a histogram.

Example - KDE

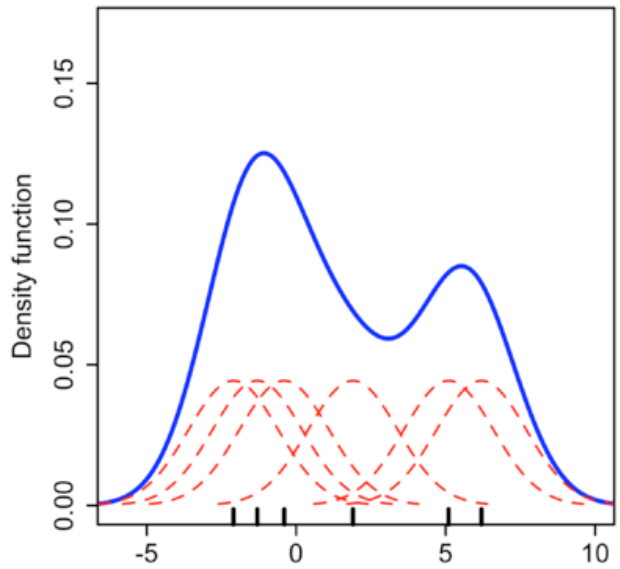

Figure 1:KDE for $x_1 = -2.1$, $x_2 = -1.3$, $x_3 = -0.4$, $x_4 = 1.9$, $x_5 = 5.1$, $x_6 = 6.2$

Sampling

Objectives