LECTURE 16 - BIASED/UNBIASED ESTIMATORS AND CONJUGATE PRIORS

1. Biased and Unbiased Estimators

- Last class we introduced the concept of Maximum Likelihood Estimation
- Maximum Likelihood Estimation (MLE) is a method of estimating parameters by maximizing the data likelihood with respect to those parameters
- Suppose we have assume a Gaussian data likelihood. What is the MLE solution for the mean?
- The MLE solution for the variance is
- A point estimate (e.g., the MLE estimate of the mean or variance) is said to be *unbiased* if the expected value of that estimator equals the "true" underlying value.
- In other words, if you have data generated from a Gaussian distribution and you used that data to compute the MLE of the mean of the Gaussian distribution. If the expected value of the MLE of the mean equals the true mean of the Gaussian, then the MLE solution is *unbiased*. If the expected value does not equal the true value, then the point estimate is *biased*.

1.1. MLE of Mean of Gaussian.

– Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ be samples from a multi-variance Normal distribution with known covariance matrix and an unknown mean. Given this data, obtain the ML estimate of the mean vector.

$$p(\mathbf{x}_k|\mu) = \frac{1}{(2\pi)^{\frac{l}{2}}\,|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}_k-\mu)^T\Sigma^{-1}(\mathbf{x}_k-\mu)\right) \tag{1}$$

– We can define our likelihood given the N data points. We are assuming these data points are drawn independently but from an identical distribution (i.i.d.):

$$\prod_{n=1}^{N} p(\mathbf{x}_n|\mu) = \prod_{n=1}^{N} \frac{1}{(2\pi)^{\frac{l}{2}}\,|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}_n-\mu)^T\Sigma^{-1}(\mathbf{x}_n-\mu)\right) \tag{2}$$

– We can apply our "trick" to simplify

$$\mathscr{L} = \ln \prod_{n=1}^{N} p(\mathbf{x}_n|\mu) = \ln \prod_{n=1}^{N} \frac{1}{(2\pi)^{\frac{l}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \mu)^T \Sigma^{-1}(\mathbf{x}_n - \mu)\right) \tag{3}$$

$$= \sum_{n=1}^{N} \ln \frac{1}{(2\pi)^{\frac{l}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \mu)^T \Sigma^{-1}(\mathbf{x}_n - \mu)\right) \tag{4}$$

$$= \sum_{n=1}^{N} \left(\ln \frac{1}{(2\pi)^{\frac{l}{2}} |\Sigma|^{\frac{1}{2}}} + \left(-\frac{1}{2}(\mathbf{x}_n - \mu)^T \Sigma^{-1}(\mathbf{x}_n - \mu)\right)\right) \tag{5}$$

$$= -N \ln(2\pi)^{\frac{l}{2}} |\Sigma|^{\frac{1}{2}} + \sum_{n=1}^{N} \left(-\frac{1}{2}(\mathbf{x}_n - \mu)^T \Sigma^{-1}(\mathbf{x}_n - \mu)\right) \tag{6}$$

– Now, lets maximize:

$$\frac{\partial \mathscr{L}}{\partial \mu} = \frac{\partial}{\partial \mu}\left[-N \ln(2\pi)^{\frac{l}{2}} |\Sigma|^{\frac{1}{2}} + \sum_{n=1}^{N} \left(-\frac{1}{2}(\mathbf{x}_n - \mu)^T \Sigma^{-1}(\mathbf{x}_n - \mu)\right)\right] = 0 \tag{7}$$

$$\rightarrow \sum_{n=1}^{N} \Sigma^{-1}(\mathbf{x}_n - \mu) = 0 \tag{8}$$

$$\rightarrow \sum_{n=1}^{N} \Sigma^{-1}\mathbf{x}_n = \sum_{n=1}^{N} \Sigma^{-1}\mu \tag{9}$$

$$\rightarrow \Sigma^{-1} \sum_{n=1}^{N} \mathbf{x}_n = \Sigma^{-1}\mu N \tag{10}$$

$$\rightarrow \sum_{n=1}^{N} \mathbf{x}_n = \mu N \tag{11}$$

$$\rightarrow \frac{\sum_{n=1}^{N} \mathbf{x}_n}{N} = \mu \tag{12}$$

(13)

– So, the ML estimate of μ is the sample mean!

1.2. Is MLE of Mean of Gaussian a Biased or Unbiased Estimator?

– An "unbiased" estimator means that the expected value of the estimator equals the true value.

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] = \frac{1}{N} N \mathbb{E}[x] = \mathbb{E}[x] = \mu \tag{14}$$

– First equality makes use of the fact that the expected value of any (i.i.d.) sample is the same.

– We showed the last equality in the last class.

1.3. **Is MLE of Variance of Gaussian a Biased or Unbiased Estimator?**

– The ML estimate of the variance of a Gaussian given N samples and a known mean is:

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2 \tag{15}$$

– The expected value of this estimator is:

$$\mathbb{E}[\sigma^2_{MLE}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right] = \frac{N-1}{N}\sigma^2 \tag{16}$$

– So, the $\mathbb{E}[\sigma^2_{MLE}] \neq \sigma^2$. Thus, the MLE of the variance of a Gaussian is a *biased* estimate.

– The *unbiased* estimate of the variance turns out to be:

$$\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu)^2 \tag{17}$$

2. Method of Maximum A Posteriori (MAP)

- Bayes Rule: $p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$
- Lets look at this in terms of binary variables, e.g., Flipping a coin: $X = 1$ is "heads", $X = 2$ is "tails"
- Let μ be the probability of heads. If we know μ, then: $P(x = 1|\mu) = \mu$ and $P(x = 0|\mu) = 1 - \mu$

$$P(x|\mu) = \mu^x(1-\mu)^{1-x} = \begin{cases} \mu & \text{if } x = 1 \\ 1-\mu & \text{if } x = 0 \end{cases} \tag{18}$$

- This is called the *Bernoulli* distribution

$$E[x] = \mu \tag{19}$$

$$E\left[(x-\mu)^2\right] = \mu(1-\mu) \tag{20}$$

- So, suppose we conducted many Bernoulli trials (e.g., flip a coin) and we want to estimate μ
- **Method: Maximum Likelihood**

$$p(\mathscr{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) \tag{21}$$

$$= \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} \tag{22}$$

- Maximize : (*What trick should we use?*)

$$\mathscr{L} = \sum_{n=1}^{N} x_n \ln \mu + (1 - x_n) \ln(1 - \mu) \tag{23}$$

$$\frac{\partial \mathscr{L}}{\partial \mu} = 0 = \frac{1}{\mu} \sum_{n=1}^{N} x_n - \frac{1}{1-\mu} \sum_{n=1}^{N} (1 - x_n) \tag{24}$$

$$0 = \frac{(1-\mu) \sum_{n=1}^{N} x_n - \mu \sum_{n=1}^{N} (1 - x_n)}{\mu(1-\mu)} \tag{25}$$

$$0 = \sum_{n=1}^{N} x_n - \mu \sum_{n=1}^{N} x_n - \mu \sum_{n=1}^{N} 1 + \mu \sum_{n=1}^{N} x_n \tag{26}$$

$$0 = \sum_{n=1}^{N} x_n - \mu N \tag{27}$$

$$\mu = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{m}{N} \tag{28}$$

where m is the number of successful trials.

- So, if we flip a coin 1 time and get heads, then $\mu = 1$ and probability of getting tails is 0. *Would you believe that? We need a prior!*
- Look at several independent trials. Consider N = 3 and m = 2 (N is number of trials, m is number of successes) and look at all ways to get 2 H and 1 T:
  - H H T $\rightarrow \mu\mu(1-\mu) = \mu^2(1-\mu)$
  - H T H $\rightarrow \mu(1-\mu)\mu = \mu^2(1-\mu)$
  - T H H $\rightarrow (1-\mu)\mu\mu = \mu^2(1-\mu)$
- $\begin{pmatrix} 3 \\ 2 \end{pmatrix} \mu^2(1-\mu) \rightarrow \begin{pmatrix} N \\ m \end{pmatrix} \mu^m (1-\mu)^{N-m} = \frac{N!}{(N-m)!m!} \mu^m (1-\mu)^{N-m}$
- This is the Binomial Distribution, gives the probability of m observations of $x = 1$ out of N independent trails
- So, what we saw is that we need a prior. We want to incorporate our prior belief. Let us place a prior on μ

$$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \tag{29}$$

$$E[\mu] = \frac{a}{a+b} \tag{30}$$

$$Var[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \tag{31}$$

- Note: $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$ and when x is an integer, then it simplifys to $x!$

- Calculation of the posterior, Take $N = m + l$ observations:

$$p(\mu|m,l,a,b) \propto Bin(m,l|\mu)Beta(\mu|a,b) \tag{32}$$

$$\propto \mu^m(1-\mu)^l\mu^{a-1}(1-\mu)^{b-1} \tag{33}$$

$$= \mu^{m+a-1}(1-\mu)^{l+b-1} \tag{34}$$

- What does this look like? Beta: $a \leftarrow m + a$, $b \leftarrow l + b$
- So, whats to posterior?

$$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1} \tag{35}$$

- **Conjugate Prior Relationship:** When the posterior is the same form as the prior
- Now we can maximize the (log of the) posterior:

$$\max_\mu (m+a-1)\ln\mu + (l+b-1)\ln(1-\mu) \tag{36}$$

$$\frac{\partial \mathscr{L}}{\partial \mu} = 0 = \frac{m+a-1}{\mu} - \frac{l+b-1}{1-\mu} \tag{37}$$

$$= (1-\mu)(m+a-1) - \mu(l+b-1) \tag{38}$$

$$= (m+a-1) - \mu(m+a-1) - \mu(l+b-1) \tag{39}$$

$$\mu = \frac{m+a-1}{m+a+l+b-2} \tag{40}$$

- This is the MAP solution. *So, what happens now when you flip one heads, two heads, etc.?*
- Discuss online updating of the prior. Eventually the data takes over the prior.