CS 5135/6035 Learning Probabilistic Models

Lecture 7: Parameter estimation, Maximum Likelihood Estimation

Gowtham Atluri

September 13, 2018

Reading material

- In Jae Myung, Tutorial on maximum likelihood estimation: http://times.cs.uiuc.edu/course/410/note/mle.pdf
- Maximum Likelihood Estimates MIT Course 18.05: https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading10b.pdf
- Course CSC321 at Univ. of Toronto: http://www.cs.toronto.edu/~rgrosse/csc321/probabilistic_models.pdf

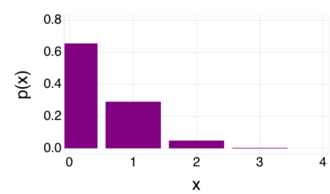
Background on parameter estimation

- Probabilistic inference vs. Parameter Estimation
- Max. Likelihood vs. Bayesian
- Desirable properties of estimators

Probabilistic Inference

- Involves computation of probabilities for events, given a model family and choices for the parameters.

E.g.: 10% of a large lot of apples are damaged. If four apples are randomly sampled from the lot, find the probability that at least one apple in the sample of four is defective. $p(x \geq 1)$?



- Given parameter a, the probability for each state of x can be computed using Binomial distribution

$$p(x) = \binom{n}{x} a^x (1-a)^{n-x}, \quad x = 0, 1, 2, \ldots n$$

Source of probability distributions?

E.g.: 10% of a large lot of apples are damaged.

Where do we get this information from?

- a is not readily available to us.
- Samples/data is available to us.
- a needs to be estimated from the data.

In general, we need to determine both the probability model and the parameters.

- We have some understanding of how to choose a suitable probability model for a given situation
- We need to estimate parameters using the collected data

Parameter Estimation

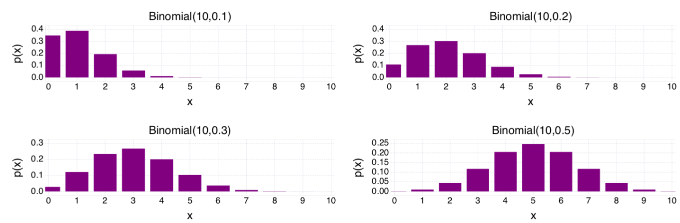
- Invovles estimation of parameters given a parametric model and observed data drawn from it.

Parameter Estimation

- Invovles estimation of parameters given a parametric model and observed data drawn from it.
- E.g.: 20 apples were inspected and 3 apples were found to be damaged.
  - We need to estimate a from this data.

Parameter Estimation

- Invovles estimation of parameters given a parametric model and observed data drawn from it.
- E.g.: 20 apples were inspected and 3 apples were found to be damaged.
  - We need to estimate a from this data.
  - Why is it non-trivial? This (3 out of 20) can be a result several Binomial distributions, which one would have generated this.



Approaches for parameter estimation - I

Maximum Likelihood Estimation (MLE)

- Parameters are assumed to be **fixed** but unknown
- ML solution seeks the solution that *best explains the dataset X*

$$\hat{\theta}_{MLE} = argmax_{\theta} p(X|\theta)$$

Approaches for parameter estimation - II

Bayesian Estimation

- Parameters are assumed to be random variables with some known a *priori* distribution $p(\theta)$
- Bayesian methods seek to estimate the posterior density $p(\theta|X)$

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

$$\hat{\theta}_{MAP} = argmax_{\theta} p(\theta|X)$$

Properties of Estimators

Consistency

An estimator is consistent if the estimate $\hat{\theta}$ it constructs is guaranteed to converge to the true parameter value θ as the quantitiy of data to which it is applied increases.

Properties of Estimators

Consistency

An estimator is consistent if the estimate $\hat{\theta}$ it constructs is guaranteed to converge to the true parameter value θ as the quantitiy of data to which it is applied increases.

Bias

The bias of an estimator η is defined as the deviation of the expectation of the estimate from the true value: $E[\hat{\theta}_{\eta}]$
When the sampling of data is viewed as a stochastic process, then the estimated parameter $\hat{\theta}_{\eta}$ cab be viewed as a continuous random variable.
When $E[\hat{\theta}_{\eta}] = \theta$ we say the estimator is unbiased.}

Properties of Estimators

Consistency

An estimator is consistent if the estimate $\hat{\theta}$ it constructs is guaranteed to converge to the true parameter value θ as the quantitiy of data to which it is applied increases.

Bias

The bias of an estimator η is defined as the deviation of the expectation of the estimate from the true value: $E[\hat{\theta}_\eta]$
When the sampling of data is viewed as a stochastic process, then the estimated parameter $\hat{\theta}_\eta$ cab be viewed as a continuous random variable.
When $E[\hat{\theta}_\eta] = \theta$ we say the estimator is unbiased.}

Variance (and efficiency)

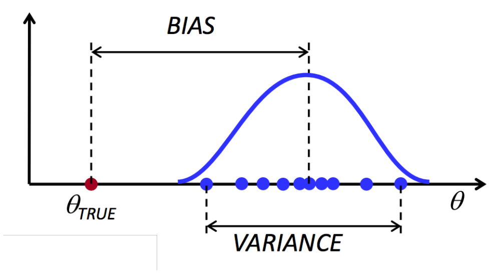
$Var[\hat{\theta}_\eta]$
All else being equal, an estimator with smaller variance is preferable to one with greater variance.

Assessing Bias in an estimator: Example

- The question is whether or not the estimator differs from the actual parameter.
- If the actual value is not different from the estimated value, we call it an unbiased estimator.
- Consider the example of computation of the mean #heads from 16 flips of a fair coin.
- Have 10 individuals do this experiment and report their observations
- For a fair coin, the expected number of heads is 8.
- Mean over 10 experiments is $\frac{1}{10}(8+8+8+8+8+8+8+8+8+8) = 8$
- Let the observations from 10 individuals be $\{11, 7, 7, 6, 6, 9, 12, 6, 10, 7\}$
- The estimate computed from these observations is 8.1.

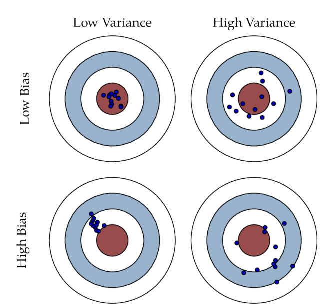
Bias vs. Variance

- **Bias**: How close is the estimate to the true value (on average)?
- **Variance**: How much does it change for different datasets?



Bias-Variance Tradeoff

- In most cases, you can only decrease one of them at the expense of the other
- Due to conflict in trying to simultaneously minimize these two sources of error
  1. Error due to wrong assumptions in the learning algorithm (misses regularities in the data - underfitting)
  2. Error from sensitivity to small fluctuations (noise) in the data (models noise in the data - overfitting)



Maximum Likelihood Estimation

- I.I.D assumption
- Likelihood
- Log-likelihood
- Maximization

Running Examples

Example 1

In a coin toss experiment where a coin was flipped 10 times, the results were H T T H H H T H H T (6 heads and 4 tails). What is the probability of seeing a head using the coin that generated these observations

Example 2

The temperatures, in Celsius, in Minneapolis during the first week of March 2018 are observed as $(-2.5, -9.9, -12.1, -8.9, -6.0, -4.8, 2.4)$
What is the distribution from which this data was generated (assuming it was Gaussian)?

Random Sampling: Independent and Identically Distributed

- Sampling depends on several items
  - Distribution (along with the parameters)
  - Sample size, n
  - Method of sampling (with or without replacement)
- $x_1, x_2, \ldots, x_n$ form a random sample of size n if:
  1. x_i's are *independently* sampled
  2. Ever x_i is drawn from the same probability distribution, i.e., *identically distributed*
- If a random sample satisfies the above two properties, we say x_i's are i.i.d.

Joint probability for observations/samples/data

- The probability density function (pdf) of a r.v. x, conditioned on the set of parameters θ, is denoted as $f(x|\theta)$.
  - This function identifies the data generating process that underlies an observed sample of data
- For a variable x, a set of i.i.d. observations $D = (x_1, \ldots, x_n)$ are drawn using a pdf $f(x|\theta)$
- The joint density of n i.i.d. observations from this process is

$$p(x_1, \ldots, x_n|\theta)$$

- As each of the observations are *independent* and *identically distributed*

$$p(x_1, \ldots, x_n|\theta) = p(x_1|\theta)p(x_2|\theta)\ldots p(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Likelihood function $L(\theta|D)$

- This joint density is the **likelihood function**, defined as a function of the unknown parameter θ

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = L(\theta|x)$$

- Note that we write the joint density of observations as $p(D|\theta)$, whereas the likelihood function, is written as $L(\theta|D)$
- Likelihood function is also denoted as $L(\theta)$ for simplicity

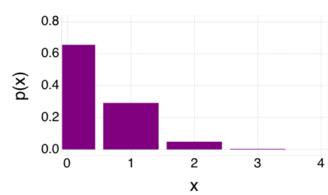
PDF vs. Likelihood function

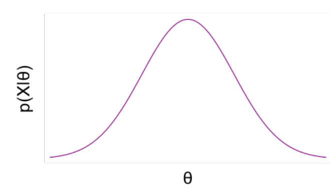
Probability density function $p(x)$

- Function of state of r.v. x
- Normalized $\sum_x p(x) = 1$

Likelihood function $L(\theta|x)$

- Function of unknown param. θ
- Unnormalized

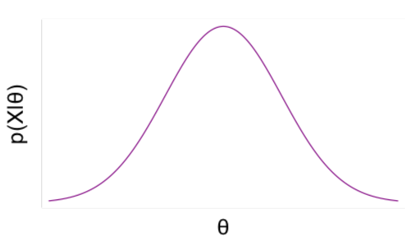




Maximum Likelihood Estimation (MLE)

- MLE solution seeks the solution that *best explains the dataset X*

$$\hat{\theta}_{MLE} = argmax_\theta p(D|\theta) = argmax_\theta L(\theta)$$



Likelihood: Example I

In a coin toss experiment where a coin was flipped 10 times, 6 heads and 4 tails were observed. What is the likelihood function $L(\theta)$?

Let θ be the probability for seeing heads in a coin toss.

$$\begin{aligned} L(\theta|D) &= p(6H, 4T|\theta) \\ &= p(H|\theta)^6 p(T|\theta)^4 \\ &= \theta^6(1-\theta)^4 \end{aligned}$$

Likelihood: Example II

The temperatures, in Celsius, in Minneapolis during the first week of March 2018 are observed as

$$-2.5 \quad -9.9 \quad -12.1 \quad -8.9 \quad -6.0 \quad -4.8 \quad 2.4$$

If this data follows a Gaussian distribution, what is the likelihood function?

Let the 7 values be denoted as $D = (x_1, \ldots, x_7)$ Given that this data follows Gaussian distribution, let the parameters be mean μ and var. σ^2.

$$L(\theta|D) = p((x_1, \ldots, x_7)|\mu, \sigma^2) = \prod_{i=1}^{7} \mathcal{N}(x_i|\mu, \sigma^2)$$

$$L(\theta|D) = (\frac{1}{\sigma\sqrt{2\pi}})^7 e^{-\sum_{i=1}^{7}(x_i-\mu)^2/2\sigma^2}$$

Log-likelihood

- Likelihood $L(\theta|D)$ for n observations is a product of n probabilities
  - where each $0 \leq p(x_i|\theta) \leq 1$
- As a result, Likelihood function will taken on extremely small values
  - $L(\theta) = \theta^{NH}(1-\theta)^{NT}$ $\qquad L(0.5) = 0.5^{100} \approx 7.9 \times 10^{-31}$

Log-likelihood

- Likelihood $L(\theta|D)$ for n observations is a product of n probabilities
  - where each $0 \leq p(x_i|\theta) \leq 1$
- As a result, Likelihood function will taken on extremely small values
  - $L(\theta) = \theta^{NH}(1-\theta)^{NT}$ $\qquad L(0.5) = 0.5^{100} \approx 7.9 \times 10^{-31}$
- **log-likelihood function** helps to avoid numerical underflow

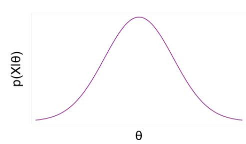
$$\ell(\theta) = logL(\theta) \qquad \ell(0.5) = log(0.5^{100}) = -69.31$$

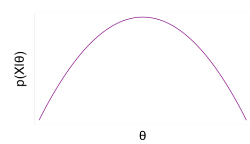
Log-likelihood

- Likelihood $L(\theta|D)$ for n observations is a product of n probabilities
  - where each $0 \leq p(x_i|\theta) \leq 1$
- As a result, Likelihood function will taken on extremely small values
  - $L(\theta) = \theta^{NH}(1-\theta)^{NT}$ $\qquad L(0.5) = 0.5^{100} \approx 7.9 \times 10^{-31}$
- **log-likelihood function** helps to avoid numerical underflow

$$\ell(\theta) = logL(\theta) \qquad \ell(0.5) = log(0.5^{100}) = -69.31$$

- Logarithm is a monotonic function, so the logarithm of a function achieves its maximum value at the same points as the function itself
- As we will see, it is also easier to differentiate for sum of terms rather than product of terms





Loglikelihood examples

Example I: 6 Heads and 4 Tails

$$\ell(\theta) = \log L(\theta) = \log(\theta^6(1-\theta)^4) = 6\log\theta + 4\log(1-\theta)$$

Example II: Temp in Minneapolis $(-2.5, -9.9, -12.1, -8.9, -6.0, -4.8, 2.4)$

$$\ell(\theta) = \log L(\theta) = \log((\frac{1}{\sigma\sqrt{2\pi}})^7 e^{-\sum_{i=1}^{7}(x_i-\mu)^2/2\sigma^2})$$

$$= -7\log(\sqrt{2\pi}) - 7\log(\sigma) - \sum_{i=1}^{7}\frac{(x_i-\mu)^2}{2\sigma^2}$$

Finding θ where Log-Likelihood is maximum - Approach I

Compute derivative and solve for params.

If possible, compute the first derivative of the log-likelihood and set it to 0. Solve for parameters.
- Compute partial derivatives when there are multiple unknown params.

Example 1: 6 Heads and 4 Tails

$$\ell(\theta) = 6\log\theta + 4\log(1-\theta)$$
$$\frac{d}{d\theta}\ell(\theta) = 6\frac{1}{\theta} - \frac{4}{1-\theta} = 0$$
$$\implies 6(1-\theta) = 4\theta$$
$$\implies \hat{\theta} = 0.6$$

Learned distribution: Example 1

In a coin toss experiment where a coin was flipped 10 times, 6 heads and 4 tails were observed. What is the probability of seeing a head using the coin that generated these observations?

```
d = Bernoulli(0.6);
xa = collect(0:1);
px = pdf.(d,xa);
myplot = plot(x=xa,y=px,Geom.bar,Coord.Cartesian(xmin=0, xmax=
              Guide.xticks(ticks=collect(0:1)), Guide.ylabel('
draw(PNG("./figs/bernouli_dist.png", 10inch, 2inch), myplot);
```

Finding θ where Log-Likelihood is maximum - Approach I

Example II: Temp in Minneapolis $(-2.5, -9.9, -12.1, -8.9, -6.0, -4.8, 2.4)$

$$\ell(\theta) = \log L(\theta) = -7\log(\sqrt{2\pi}) - 7\log(\sigma) - \sum_{i=1}^{7} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Since $\ell(\theta)$ has two params μ and σ we will use partial derivatives.

$$\frac{\partial \ell(\mu,\sigma)}{\partial \mu} = \sum_{i=1}^{7} \frac{x_i-\mu}{\sigma^2} = 0 \implies \sum_{i=1}^{7} x_i = 7\mu \implies \hat{\mu} = \frac{\sum_{i=1}^{7} x_i}{7} = -5.97$$

$$\frac{\partial \ell(\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{7} \frac{(x_i-\mu)^2}{\sigma^3} = 0 \implies \hat{\sigma}^2 = \frac{\sum_{i=1}^{7}(x_i-\mu)^2}{n} = 20.72$$

Learned distribution: Example 2

The temperatures, in Celsius, in Minneapolis during the first week of March 2018 are observed as $(-2.5, -9.9, -12.1, -8.9, -6.0, -4.8, 2.4)$ What is the distribution from which this data was generated (assuming it was Gaussian)?

```
d = Normal(-5.97,sqrt(20.72));
xa = collect(-40:0.02:40);
px = pdf.(d,xa);
myplot = plot(x=xa,y=px,Geom.bar,Coord.Cartesian(xmin=-40, xmax=40)
             Guide.ylabel("p(x)"),white_panel);
draw(PNG("./figs/normal_dist.png", 10inch, 2inch), myplot);
```

Finding θ where Log-Likelihood is maximum - Approach II

- When approach I is not possible (particularly when the model involves many parameters and its PDF is highly non-linear), use gradient descent approach.
  - Use negative log-likelihood (also referred to as a cost function)
  - Randomly initialize and then incrementally update our weights by calculating the slope of our objective function
  - When applying the cost function, we want to continue updating our weights until the slope of the gradient gets as close to zero as possible.