College of Media, Communication and Information
UNIVERSITY OF COLORADO BOULDER

# Maximum Likelihood Estimation

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

MARCH 7, 2017

## Why MLE?

- Before: Distribution + Parameter $\longrightarrow x$
- Now: $x$ + Distribution $\longrightarrow$ Parameter
- (Much more realistic)
- But: Says nothing about how good a fit a distribution is

## Likelihood

- Likelihood is $p(x; \theta)$
- We want estimate of $\theta$ that best explains data we seen
- I.e., Maximum Likelihood Estimate (MLE)

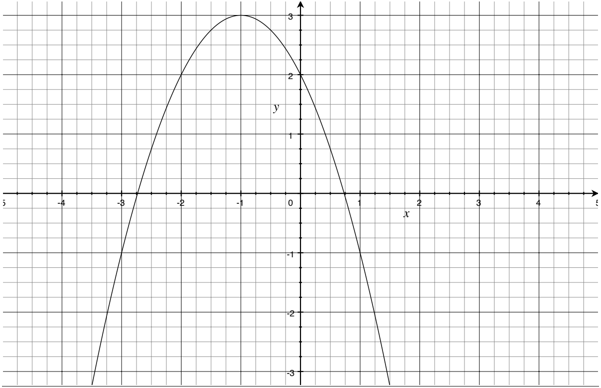
## Likelihood

- The *likelihood function* refers to the PMF (discrete) or PDF (continuous).
- For discrete distributions, the likelihood of $x$ is $P(X = x)$.
- For continuous distributions, the likelihood of $x$ is the density $f(x)$.
- We will often refer to likelihood rather than probability/mass/density so that the term applies to either scenario.

## Optimizing Unconstrained Functions

Suppose we wanted to optimize

$$\ell = x^2 - 2x + 2 \tag{1}$$

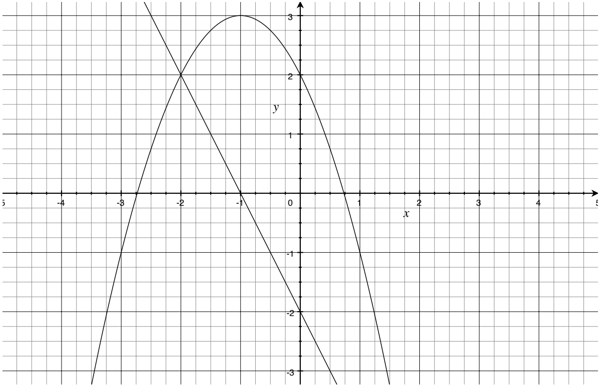


## Optimizing Unconstrained Functions

Suppose we wanted to optimize

$$\ell = x^2 - 2x + 2 \quad (1)$$

$$\frac{\partial \ell}{\partial x} = -2x - 2 \quad (2)$$



## Optimizing Unconstrained Functions

$$\frac{\partial \ell}{\partial x} = 0 \tag{3}$$

$$-2x - 2 = 0 \tag{4}$$

$$x = -1 \tag{5}$$

(Should also check that second derivative is negative)

**Theorem: Lagrange Multiplier Method**

Given functions $f(x_1, \dots x_n)$ and $g(x_1, \dots x_n)$, the critical points of $f$ restricted to the set $g = 0$ are solutions to equations:

$$\begin{aligned} \frac{\partial f}{\partial x_i}(x_1, \dots x_n) = & \lambda \frac{\partial g}{\partial x_i}(x_1, \dots x_n) \quad \forall i \\ g(x_1, \dots x_n) = & 0 \end{aligned}$$

This is $n+1$ equations in the $n+1$ variables $x_1, \dots x_n, \lambda$.

## Lagrange Example

Maximize $\ell(x, y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

## Lagrange Example

Maximize $\ell(x,y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

$$\frac{\partial \ell}{\partial x} = \frac{1}{2}\sqrt{\frac{y}{x}} \quad \frac{\partial g}{\partial x} = 20$$
$$\frac{\partial \ell}{\partial y} = \frac{1}{2}\sqrt{\frac{x}{y}} \quad \frac{\partial g}{\partial y} = 10$$

## Lagrange Example

Maximize $\ell(x,y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

$$\frac{\partial \ell}{\partial x} = \frac{1}{2}\sqrt{\frac{y}{x}} \quad \frac{\partial g}{\partial x} = 20$$
$$\frac{\partial \ell}{\partial y} = \frac{1}{2}\sqrt{\frac{x}{y}} \quad \frac{\partial g}{\partial y} = 10$$

- Create new systems of equations

## Lagrange Example

Maximize $\ell(x,y)=\sqrt{xy}$ subject to the constraint $20x+10y=200$.

- Compute derivatives

$$\frac{\partial \ell}{\partial x}=\frac{1}{2}\sqrt{\frac{y}{x}} \quad \frac{\partial g}{\partial x}=20$$
$$\frac{\partial \ell}{\partial y}=\frac{1}{2}\sqrt{\frac{x}{y}} \quad \frac{\partial g}{\partial y}=10$$

- Create new systems of equations

$$\frac{1}{2}\sqrt{\frac{y}{x}}=20\lambda$$
$$\frac{1}{2}\sqrt{\frac{x}{y}}=10\lambda$$
$$20x+10y=200$$

## Lagrange Example

- Dividing the first equation by the second gives us

$$\frac{y}{x} = 2 \tag{6}$$

- which means $y = 2x$, plugging this into the constraint equation gives:

$$20x + 10(2x) = 200$$
$$x = 5 \Rightarrow y = 10$$

College of Media, Communication and Information
UNIVERSITY OF COLORADO BOULDER

# Maximum Likelihood Estimation

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

MARCH 7, 2017

## Continuous Distribution: Gaussian

- Recall the density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

## Continuous Distribution: Gaussian

- Recall the density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

$$\ell(\mu, \sigma) \equiv -N \log \sigma - \frac{N}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 \tag{2}$$

## Continuous Distribution: Gaussian

- Recall the density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

$$\ell(\mu, \sigma) \equiv -N \log \sigma - \frac{N}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 \tag{2}$$

## Continuous Distribution: Gaussian

- Recall the density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

$$\ell(\mu, \sigma) \equiv -N \log \sigma - \frac{N}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 \tag{2}$$

- Recall the density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

$$\ell(\mu,\sigma) \equiv -N \log \sigma - \frac{N}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2 \tag{2}$$

## MLE of Gaussian $\mu$

$$\ell(\mu,\sigma) = -N \log \sigma - \frac{N}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2 \tag{3}$$

$$\frac{\partial \ell}{\partial \mu} = 0 + \frac{1}{\sigma^2}\sum_i (x_i - \mu) \tag{4}$$

## MLE of Gaussian $\mu$

$$\ell(\mu,\sigma) = -N\log\sigma - \frac{N}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2 \tag{3}$$

$$\frac{\partial \ell}{\partial \mu} = 0 + \frac{1}{\sigma^2}\sum_i (x_i-\mu) \tag{4}$$

## MLE of Gaussian $\mu$

$$\ell(\mu,\sigma) = -N\log\sigma - \frac{N}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2 \tag{3}$$

$$\frac{\partial \ell}{\partial \mu} = 0 + \frac{1}{\sigma^2}\sum_i (x_i - \mu) \tag{4}$$

## MLE of Gaussian $\mu$

$$\ell(\mu,\sigma) = -N \log \sigma - \frac{N}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2 \tag{3}$$

$$\frac{\partial \ell}{\partial \mu} = 0 + \frac{1}{\sigma^2}\sum_i (x_i - \mu) \tag{4}$$

Solve for $\mu$:

$$0 = \frac{1}{\sigma^2}\sum_i (x_i - \mu) \tag{5}$$

$$0 = \sum_i x_i - N\mu \tag{6}$$

$$\mu = \frac{\sum_i x_i}{N} \tag{7}$$

## MLE of Gaussian $\mu$

$$\ell(\mu,\sigma) = -N\log\sigma - \frac{N}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2 \tag{3}$$

$$\frac{\partial \ell}{\partial \mu} = 0 + \frac{1}{\sigma^2}\sum_i (x_i-\mu) \tag{4}$$

Solve for $\mu$:

$$0 = \frac{1}{\sigma^2}\sum_i (x_i-\mu) \tag{5}$$

$$0 = \sum_i x_i - N\mu \tag{6}$$

$$\mu = \frac{\sum_i x_i}{N} \tag{7}$$

Consistent with what we said before

## MLE of Gaussian $\sigma$

$$\ell(\mu,\sigma) = -N\log\sigma - \frac{N}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2 \tag{8}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{N}{\sigma} + 0 + \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2 \tag{9}$$

## MLE of Gaussian $\sigma$

$$\ell(\mu,\sigma) = -N\log\sigma - \frac{N}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2 \tag{8}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{N}{\sigma} + 0 + \frac{1}{\sigma^3}\sum_i (x_i - \mu)^2 \tag{9}$$

## MLE of Gaussian $\sigma$

$$\ell(\mu,\sigma) = -N \log \sigma - \frac{N}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2 \tag{8}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{N}{\sigma} + 0 + \frac{1}{\sigma^3}\sum_i (x_i - \mu)^2 \tag{9}$$

## MLE of Gaussian $\sigma$

$$\ell(\mu,\sigma) = -N \log \sigma - \frac{N}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 \tag{8}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{N}{\sigma} + 0 + \frac{1}{\sigma^3} \sum_i (x_i - \mu)^2 \tag{9}$$

## MLE of Gaussian $\sigma$

$$\ell(\mu,\sigma) = -N\log\sigma - \frac{N}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2 \tag{8}$$

$$\frac{\partial\ell}{\partial\sigma} = -\frac{N}{\sigma} + 0 + \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2 \tag{9}$$

Solve for $\sigma$:

$$0 = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2 \tag{10}$$

$$\frac{N}{\sigma} = \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2 \tag{11}$$

$$\sigma^2 = \frac{\sum_i (x_i-\mu)^2}{N} \tag{12}$$

## MLE of Gaussian $\sigma$

$$\ell(\mu,\sigma) = -N \log \sigma - \frac{N}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2 \tag{8}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{N}{\sigma} + 0 + \frac{1}{\sigma^3}\sum_i (x_i - \mu)^2 \tag{9}$$

Solve for $\sigma$:

$$0 = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_i (x_i - \mu)^2 \tag{10}$$

$$\frac{N}{\sigma} = \frac{1}{\sigma^3}\sum_i (x_i - \mu)^2 \tag{11}$$

$$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{N} \tag{12}$$

Consistent with what we said before

# Maximum Likelihood Estimation

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

MARCH 7, 2017

**Theorem: Lagrange Multiplier Method**

Given functions $f(x_1, \dots x_n)$ and $g(x_1, \dots x_n)$, the critical points of $f$ restricted to the set $g = 0$ are solutions to equations:

$$\begin{aligned} \frac{\partial f}{\partial x_i}(x_1, \dots x_n) =& \lambda \frac{\partial g}{\partial x_i}(x_1, \dots x_n) \quad \forall i \\ g(x_1, \dots x_n) =& 0 \end{aligned}$$

This is $n+1$ equations in the $n+1$ variables $x_1, \dots x_n, \lambda$.

## Lagrange Example

Maximize $\ell(x, y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

## Lagrange Example

Maximize $\ell(x,y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

$$\frac{\partial \ell}{\partial x} = \frac{1}{2}\sqrt{\frac{y}{x}} \quad \frac{\partial g}{\partial x} = 20$$
$$\frac{\partial \ell}{\partial y} = \frac{1}{2}\sqrt{\frac{x}{y}} \quad \frac{\partial g}{\partial y} = 10$$

## Lagrange Example

Maximize $\ell(x,y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

$$\frac{\partial \ell}{\partial x} = \frac{1}{2}\sqrt{\frac{y}{x}} \quad \frac{\partial g}{\partial x} = 20$$
$$\frac{\partial \ell}{\partial y} = \frac{1}{2}\sqrt{\frac{x}{y}} \quad \frac{\partial g}{\partial y} = 10$$

- Create new systems of equations

## Lagrange Example

Maximize $\ell(x,y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

$$\frac{\partial \ell}{\partial x} = \frac{1}{2}\sqrt{\frac{y}{x}} \quad \frac{\partial g}{\partial x} = 20$$
$$\frac{\partial \ell}{\partial y} = \frac{1}{2}\sqrt{\frac{x}{y}} \quad \frac{\partial g}{\partial y} = 10$$

- Create new systems of equations

$$\frac{1}{2}\sqrt{\frac{y}{x}} = 20\lambda$$
$$\frac{1}{2}\sqrt{\frac{x}{y}} = 10\lambda$$
$$20x + 10y = 200$$

## Lagrange Example

- Dividing the first equation by the second gives us

$$\frac{y}{x} = 2 \tag{1}$$

- which means $y = 2x$, plugging this into the constraint equation gives:

$$\begin{aligned} 20x + 10(2x) &= 200 \\ x = 5 &\Rightarrow y = 10 \end{aligned}$$

## Discrete Distribution: Multinomial

- Recall the mass function ($N$ is total number of observations, $x_i$ is the number for each cell, $\theta_i$ probability of cell)

$$p(\vec{x} \,|\, \vec{\theta}) = \frac{N!}{\prod_i x_i!} \prod \theta_i^{x_i} \tag{2}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

## Discrete Distribution: Multinomial

- Recall the mass function ($N$ is total number of observations, $x_i$ is the number for each cell, $\theta_i$ probability of cell)

$$p(\vec{x} \,|\, \vec{\theta}) = \frac{N!}{\prod_i x_i!} \prod \theta_i^{x_i} \tag{2}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

$$\ell(\vec{\theta}) \equiv \log(n!) - \sum_i \log(x_i!) + \sum_i x_i \log \theta_i \tag{3}$$

## Discrete Distribution: Multinomial

- Recall the mass function ($N$ is total number of observations, $x_i$ is the number for each cell, $\theta_i$ probability of cell)

$$p(\vec{x}\,|\,\vec{\theta}) = \frac{N!}{\prod_i x_i!} \prod \theta_i^{x_i} \tag{2}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

$$\ell(\vec{\theta}) \equiv \log(n!) - \sum_i \log(x_i!) + \sum_i x_i \log \theta_i \tag{3}$$

## Discrete Distribution: Multinomial

- Recall the mass function ($N$ is total number of observations, $x_i$ is the number for each cell, $\theta_i$ probability of cell)

$$p(\vec{x} \mid \vec{\theta}) = \frac{N!}{\prod_i x_i!} \prod \theta_i^{x_i} \tag{2}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

$$\ell(\vec{\theta}) \equiv \log(n!) - \sum_i \log(x_i!) + \sum_i x_i \log \theta_i \tag{3}$$

## Discrete Distribution: Multinomial

- Recall the mass function ($N$ is total number of observations, $x_i$ is the number for each cell, $\theta_i$ probability of cell)

$$p(\vec{x} \,|\, \vec{\theta}) = \frac{N!}{\prod_i x_i!} \prod \theta_i^{x_i} \tag{2}$$

- Taking the log makes math easier, doesn't change answer (monotonic)
- If we observe $x_1 \dots x_N$, then log likelihood is

$$\ell(\vec{\theta}) \equiv \log(n!) - \sum_i \log(x_i!) + \sum_i x_i \log \theta_i \tag{3}$$

## MLE of Multinomial $\theta$

$$\ell(\vec{\theta}) = \log(N!) - \sum_i \log(x_i!) + \sum_i x_i \log \theta_i + \lambda \left(1 - \sum_i \theta_i\right) \tag{4}$$

$$\tag{5}$$

## MLE of Multinomial $\theta$

$$\ell(\vec{\theta}) = \log(N!) - \sum_i \log(x_i!) + \sum_i x_i \log \theta_i + \lambda\left(1 - \sum_i \theta_i\right) \tag{4}$$

$$\tag{5}$$

Where did this come from? Constraint that $\vec{\theta}$ must be a distribution.

## MLE of Multinomial $\theta$

$$\ell(\vec{\theta}) = \log(N!) - \sum_i \log(x_i!) + \sum_i x_i \log\theta_i + \lambda\left(1 - \sum_i \theta_i\right) \tag{4}$$

$$\tag{5}$$

- $\frac{\partial \ell}{\partial \theta_i} = \frac{x_i}{\theta_i} - \lambda$
- $\frac{\partial \ell}{\partial \lambda} = 1 - \sum_i \theta_i$

## MLE of Multinomial $\theta$

$$\ell(\vec{\theta}) = \log(N!) - \sum_i \log(x_i!) + \sum_i x_i \log \theta_i + \lambda\left(1 - \sum_i \theta_i\right) \tag{4}$$

$$\tag{5}$$

- $\frac{\partial \ell}{\partial \theta_i} = \frac{x_i}{\theta_i} - \lambda$
- $\frac{\partial \ell}{\partial \lambda} = 1 - \sum_i \theta_i$

## MLE of Multinomial $\theta$

$$\ell(\vec{\theta}) = \log(N!) - \sum_i \log(x_i!) + \sum_i x_i \log \theta_i + \lambda \left(1 - \sum_i \theta_i\right) \tag{4}$$

$$\tag{5}$$

- $\frac{\partial \ell}{\partial \theta_i} = \frac{x_i}{\theta_i} - \lambda$
- $\frac{\partial \ell}{\partial \lambda} = 1 - \sum_i \theta_i$

## MLE of Multinomial $\theta$

$$\ell(\vec{\theta}) = \log(N!) - \sum_i \log(x_i!) + \sum_i x_i \log \theta_i + \lambda\left(1 - \sum_i \theta_i\right) \tag{4}$$

$$\tag{5}$$

- $\frac{\partial \ell}{\partial \theta_i} = \frac{x_i}{\theta_i} - \lambda$
- $\frac{\partial \ell}{\partial \lambda} = 1 - \sum_i \theta_i$

## MLE of Multinomial $\theta$

- We have system of equations

$$\theta_1 = \frac{x_1}{\lambda} \tag{6}$$

$$\vdots \quad \vdots \tag{7}$$

$$\theta_K = \frac{x_K}{\lambda} \tag{8}$$

$$\sum_i \theta_i = 1 \tag{9}$$

## MLE of Multinomial $\theta$

- We have system of equations

$$\theta_1 = \frac{x_1}{\lambda} \tag{6}$$

$$\vdots \quad \vdots \tag{7}$$

$$\theta_K = \frac{x_K}{\lambda} \tag{8}$$

$$\sum_i \theta_i = 1 \tag{9}$$

- So let's substitute the first $K$ equations into the last:

$$\sum_i \frac{x_i}{\lambda} = 1 \tag{10}$$

## MLE of Multinomial $\theta$

- We have system of equations

$$\theta_1 = \frac{x_1}{\lambda} \tag{6}$$

$$\vdots \quad \vdots \tag{7}$$

$$\theta_K = \frac{x_K}{\lambda} \tag{8}$$

$$\sum_i \theta_i = 1 \tag{9}$$

- So let's substitute the first $K$ equations into the last:

$$\sum_i \frac{x_i}{\lambda} = 1 \tag{10}$$

- So $\lambda = \sum_i x_i = N$,

## MLE of Multinomial $\theta$

- We have system of equations

$$\theta_1 = \frac{x_1}{\lambda} \tag{6}$$

$$\vdots \quad \vdots \tag{7}$$

$$\theta_K = \frac{x_K}{\lambda} \tag{8}$$

$$\sum_i \theta_i = 1 \tag{9}$$

- So let's substitute the first $K$ equations into the last:

$$\sum_i \frac{x_i}{\lambda} = 1 \tag{10}$$

- So $\lambda = \sum_i x_i = N$, and $\theta_i = \frac{x_i}{N}$

# Maximum Likelihood Estimation

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

MARCH 7, 2017

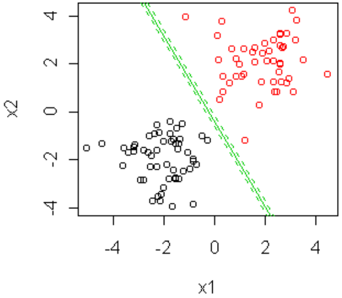
## Big Pictures

- Ran through several common examples
- For existing distributions you can (and should) look up MLE
- For new models, you can't (foreshadowing of later in class)

## Big Pictures

- Ran through several common examples
- For existing distributions you can (and should) look up MLE
- For new models, you can't (foreshadowing of later in class)
  - Classification models
  - Unsupervised models (Expectation-Maximization)
- Not always so easy

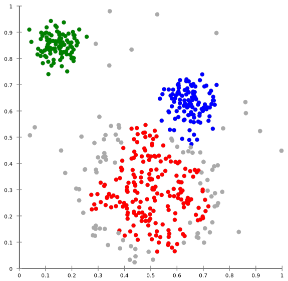
## Classification




- Classification can be viewed as $p(y|x,\theta)$
- Have $x, y$, need $\theta$
- Discovering $\theta$ is also problem of MLE

## Clustering

- Clustering can be viewed as $p(x \mid z, \theta)$
- Have $x$, need $z, \theta$
- $z$ is guessed at iteratively (Expectation)
- $\theta$ estimated to maximize likelihood (Maximization)



## Not always so easy: Bias

- An estimator is biased if $\mathbb{E}\left[\hat{\theta}\right] \neq \theta$
- We won't prove it, but the estimate for variance is biased
- Comes from estimating $\mu$, so need to "shrink" variance

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_i (x_i - \mu)^2 \tag{1}$$

## Not always so easy: Intractable Likelihoods

- Not always possible to “solve for” optimal estimator
- Use gradient optimization (we’ll see this for logistic regression)
- Use other approximations (e.g., Monte Carlo sampling)
- Whole subfield of statistics / information science