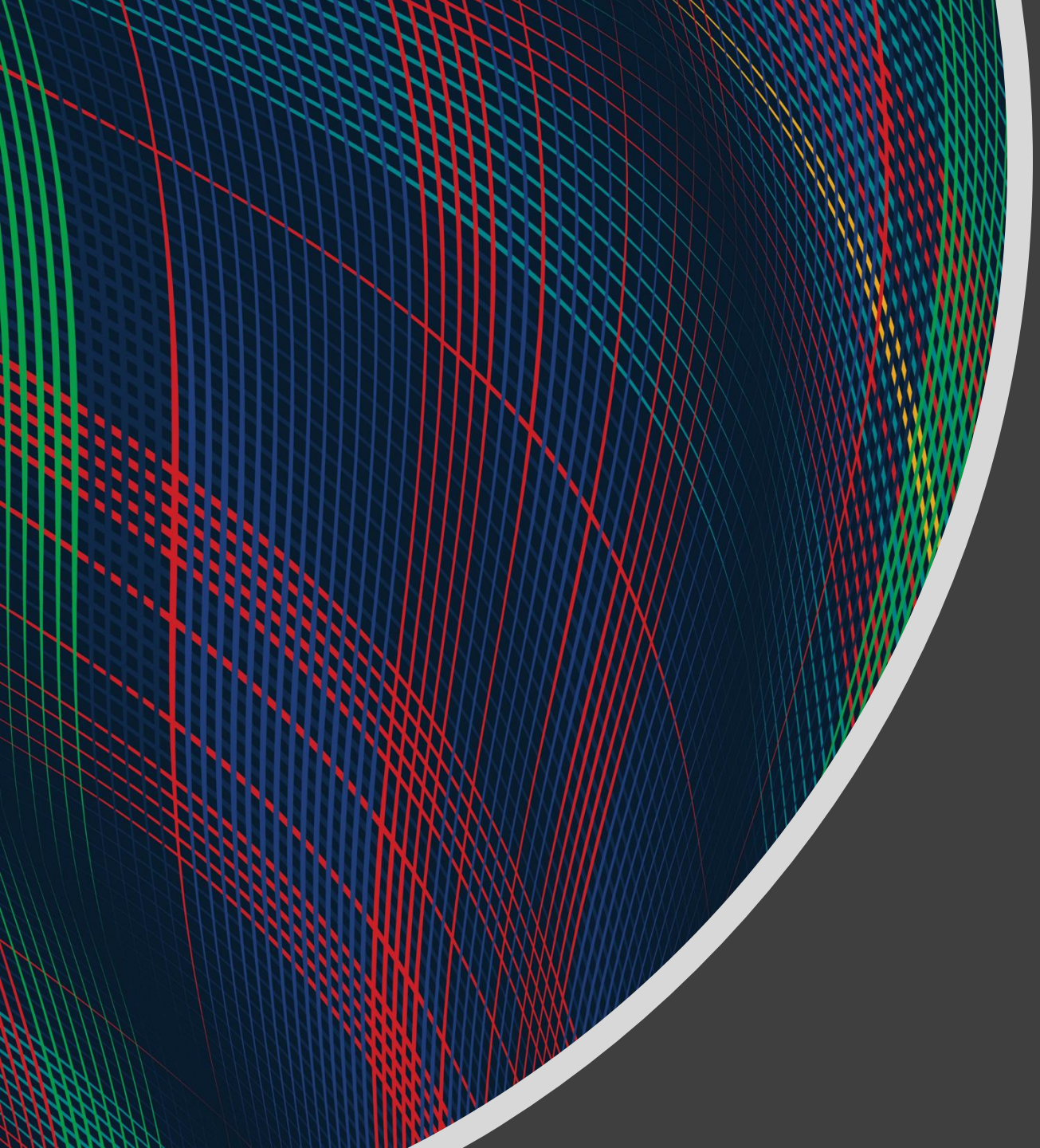


10-315 Introduction to ML

MLE and Probabilistic Formulation of Machine Learning

Instructor: Pat Virtue

Poll 1

Course feedback link on Ed

Are you finished with the course feedback form?

A. Yes
B. No
C. Still working on it

Likelihood

Pre-reading

Likelihood

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

Likelihood

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

Grades

Gaussian PDF: $p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$

Likelihood

Trick coin: comes up heads only 1/3 of the time

1 flip: H probability: $\frac{1}{3}$

2 flips: H,H probability: $\frac{1}{3} \cdot \frac{1}{3}$

3 flips: H,H,T probability: $\frac{1}{3} \cdot \frac{1}{3} \cdot \left(1 - \frac{1}{3}\right)$

But why can we just multiply these?

Likelihood and i.i.d

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

i.i.d.: Independent and identically distributed

Likelihood and Maximum Likelihood Estimation

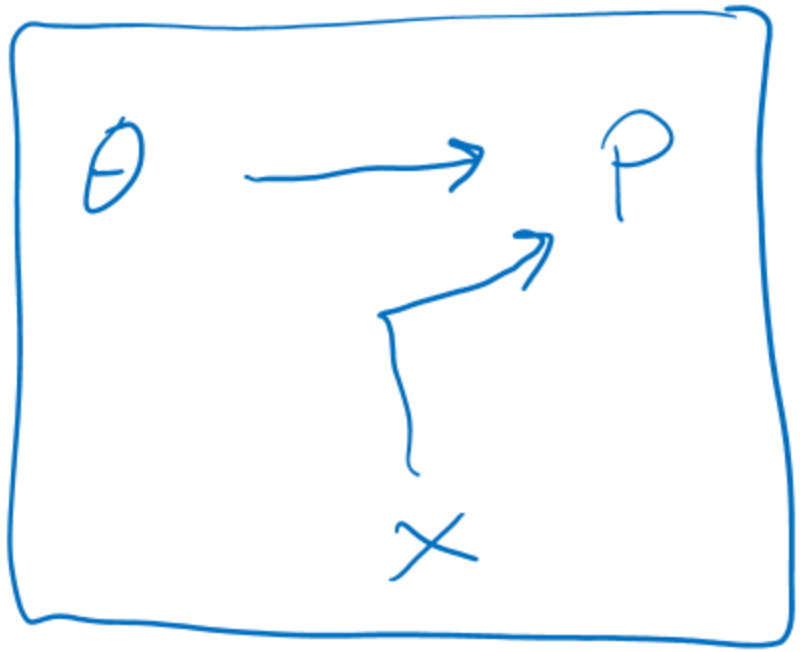
Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

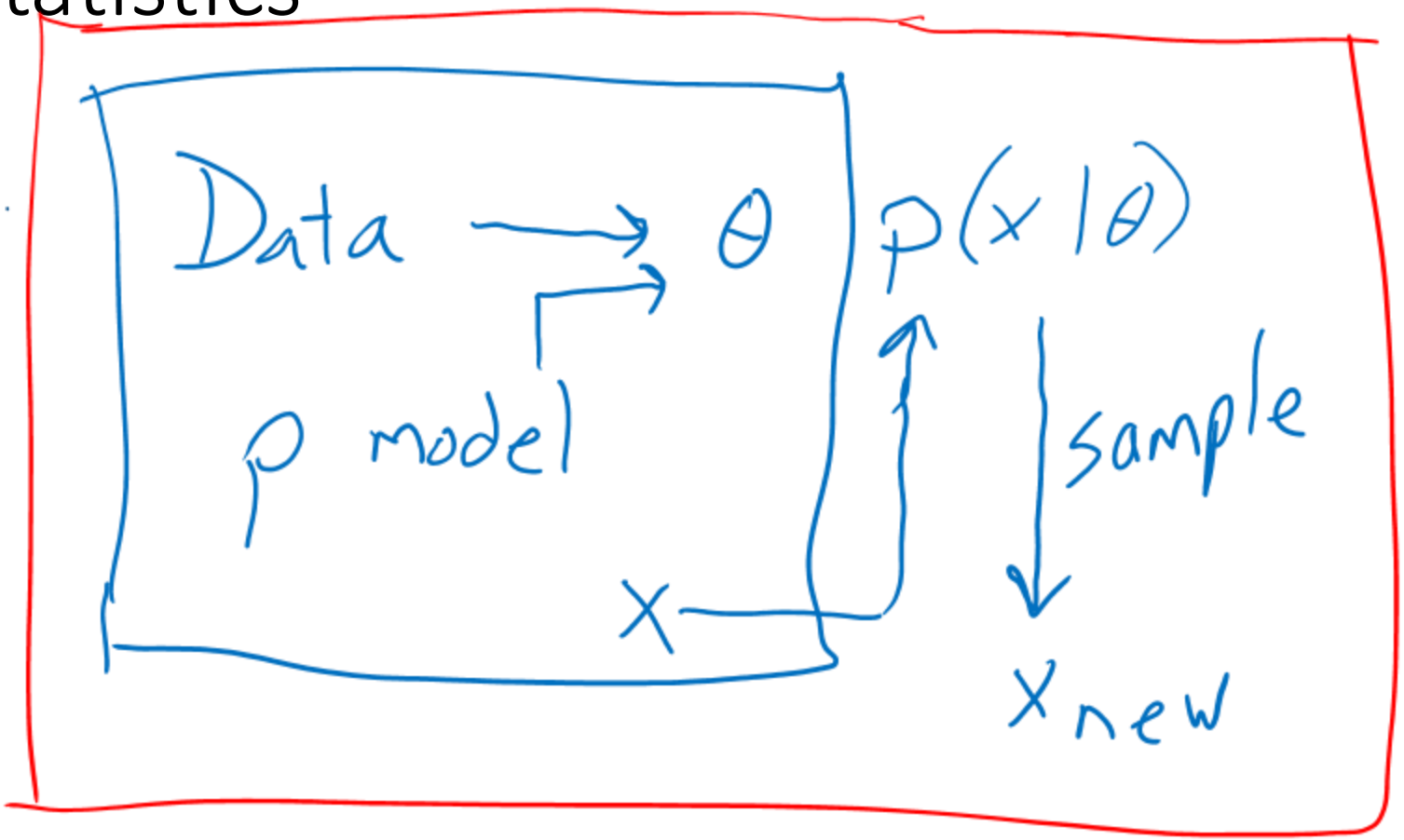
Likelihood function: The value of likelihood as we change theta

(same as likelihood, but conceptually we are considering many different values of the parameters)

Maximum Likelihood Estimation (MLE): Find the parameter value that maximizes the likelihood.

From Probability to Statistics





MLE – Logistic Regression

Bernoulli:
Categorical

Exercises

Calculate the probability of these event sequences happening

1. Coin
   a) Fair:
      {H, H, T, H}
   b) Biased, $\phi = 3/4$ heads
      {H, H, T, H}

$\frac{1}{2}\ \frac{1}{2}\ \frac{1}{2}\ \frac{1}{2}$

$\phi\ \phi\ (1-\phi)\ \phi = \frac{3 \cdot 3 \cdot 1 \cdot 3}{4\ 4\ 4\ 4} = \frac{27}{256}$

$\phi\ \phi\ \phi_T\ \phi$

2. 4-sided die with sides: A, B, C, D
   a) Fair:
      {A, B, D, D, A}
   b) Weighted, $[\phi_A, \phi_B, \phi_C, \phi_D] = [1/10,\ 2/10,\ 3/10,\ 4/10]$
      {A, B, D, D, A}

$= \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}$

$\phi_A\ \phi_B\ \phi_D\ \phi_D\ \phi_A = \frac{1}{10}\ \frac{2}{10}\ \frac{4}{10}\ \frac{4}{10}\ \frac{1}{10}$

Bernoulli Likelihood

Bernoulli distribution:

$$Y \sim Bern(\phi) \qquad p(y \mid \phi) = \begin{cases} \phi, & y = 1 \\ 1 - \phi, & y = 0 \end{cases}$$

What is the likelihood for three i.i.d. samples, given parameter ϕ:

$$\mathcal{D} = \{y^{(1)} = 1, y^{(2)} = 1, y^{(3)} = 0\}$$

$$\prod_{i=1}^{N} p(Y = y^{(i)} \mid \phi)$$

$$= \phi \cdot \phi \cdot (1 - \phi)$$

$$= \prod_{i=1}^{N} \phi^{y^{(i)}} (1-\phi)^{(1-y^{(i)})}$$

$$y^{(i)} = \left[\prod \left(y^{(i)} = 1 \right) \right]$$

$$(1-y^{(i)}) = \left[\prod \left(y^{(i)} = 0 \right) \right]$$

Bernoulli Likelihood

Bernoulli distribution:

$$Y \sim Bern(\phi) \qquad p(y \mid \phi) = \begin{cases} \phi, & y = 1 \\ 1 - \phi, & y = 0 \end{cases}$$

What is the likelihood for three i.i.d. samples, given parameter ϕ:

$$\mathcal{D} = \{y^{(1)} = 1, y^{(2)} = 1, y^{(3)} = 0\}$$

$$\prod_{i=1}^{N} p(Y = y^{(i)} \mid \phi)$$

$$= \phi \cdot \phi \cdot (1 - \phi)$$

$$\phi^1 \cdot (1-\phi)^0 \cdot \phi^1 \cdot (1-\phi)^0 \cdot \phi^0 \cdot (1-\phi)^1$$

$$= \prod_{i=1}^{N} \phi^{y^{(i)}} (1-\phi)^{(1-y^{(i)})}$$

$$y^{(i)} = \prod \left(y^{(i)} = 1\right)$$

$$(1-y^{(i)}) = \prod \left(y^{(i)} = 0\right)$$

Log Likelihood

$$\sum_{i=1}^{N} \left[y^{(i)} \log \phi + (1-y^{(i)}) \log(1-\phi) \right]$$

Estimating Parameters with Likelihood

We model the outcome of a single mysterious weighted-coin flip as a Bernoulli random variable:

$$Y \sim Bern(\phi)$$

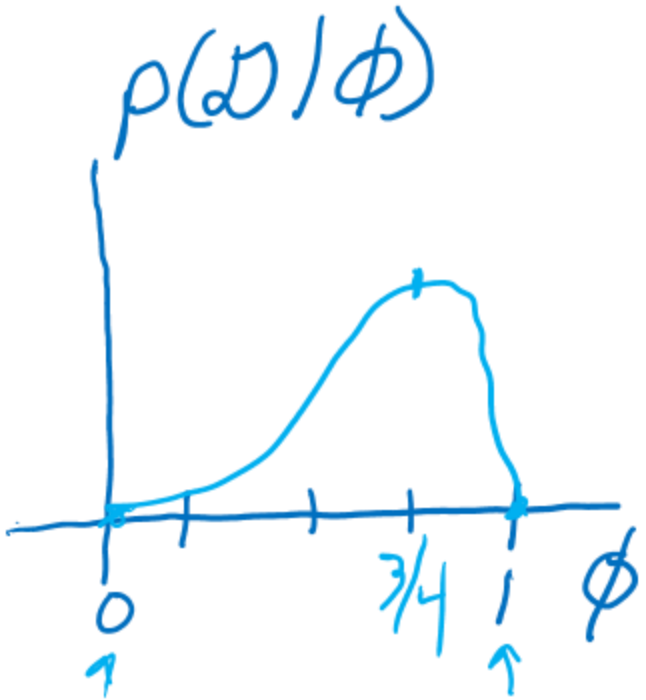
$$p(y \mid \phi) = \begin{cases} \phi, & y = 1 \ \ (heads) \\ 1 - \phi, & y = 0 \ \ (tails) \end{cases}$$

Given the ordered sequence of coin flip outcomes:

$$[1, 0, 1, 1]$$

What is the estimate of parameter $\hat{\phi}$?

$$p(D \mid \phi) = \phi \cdot \phi \cdot (1 - \phi) \cdot \phi$$
$$= \phi^3 (1 - \phi)^1$$



https://www.desmos.com/calculator/kr7m2m6cf7

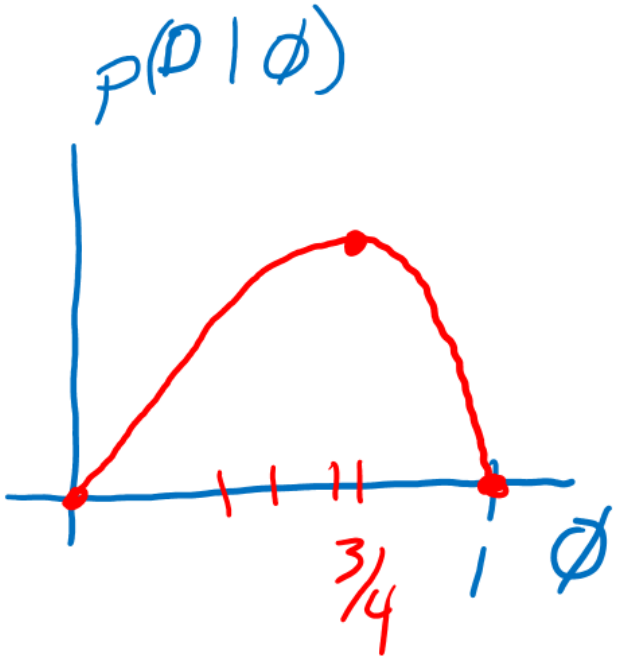
MLE as Data Increases

Given the ordered sequence of coin flip outcomes:

$$[1, 0, 1, 1]$$

$$\mathrm{p}(\mathcal{D} \mid \phi) = \prod_{i}^{N} p\left(y^{(i)} \mid \phi\right) = \phi^{N_{y=1}}(1-\phi)^{N_{y=0}}$$

What happens as we flip more coins?



https://www.desmos.com/calculator/kr7m2m6cf7

MLE for Categorical

$$Y \sim \begin{bmatrix} p(Y_1=1|\theta) \\ \vdots \\ p(Y_k=1|\theta) \\ \vdots \\ p(Y_K=1|\theta) \end{bmatrix} \leftarrow$$

$$\vec{y}^{(i)} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow$$

$$\sum_{k=1}^{K} p(Y_k=1|\theta) = 1$$

$$\mathcal{L}(\theta; D)$$

$$= \prod_{i=1}^{N} p(Y=\vec{y}^{(i)}|\theta)$$

$$= \prod_{i=1}^{N} \prod_{k=1}^{K} p(Y_k^{(i)}=1|\theta)$$

$y_k^{(i)}$

$\mathbb{I}(y_k^{(i)}=1)$

Log Likelihood

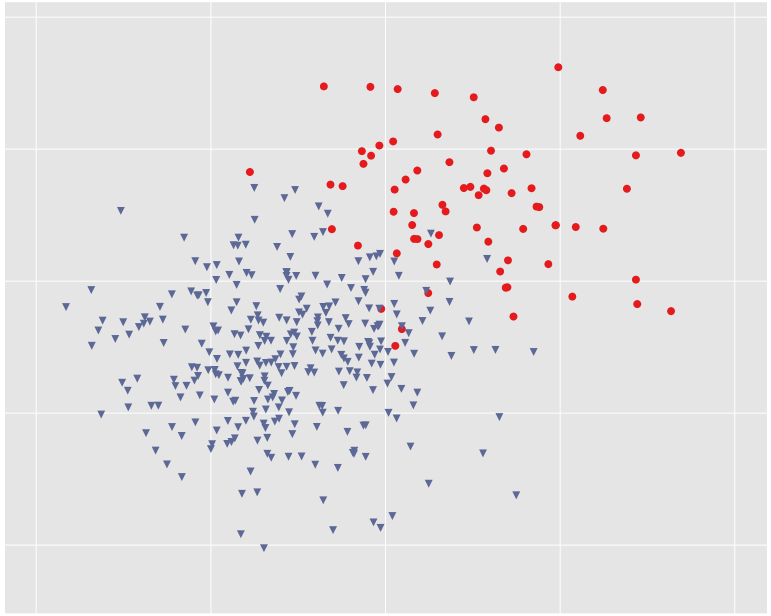
$$-\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^{(i)} \log \underbrace{p(Y_k^{(i)}=1|\theta)}_{\hat{y}_k^{(i)}}$$

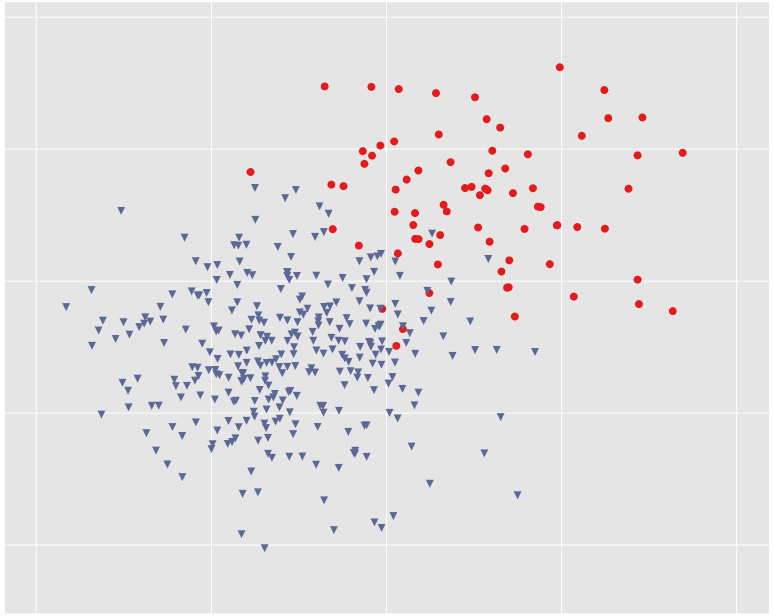
$$-\sum\sum y_k^{(i)} \log \hat{y}_k^{(i)}$$

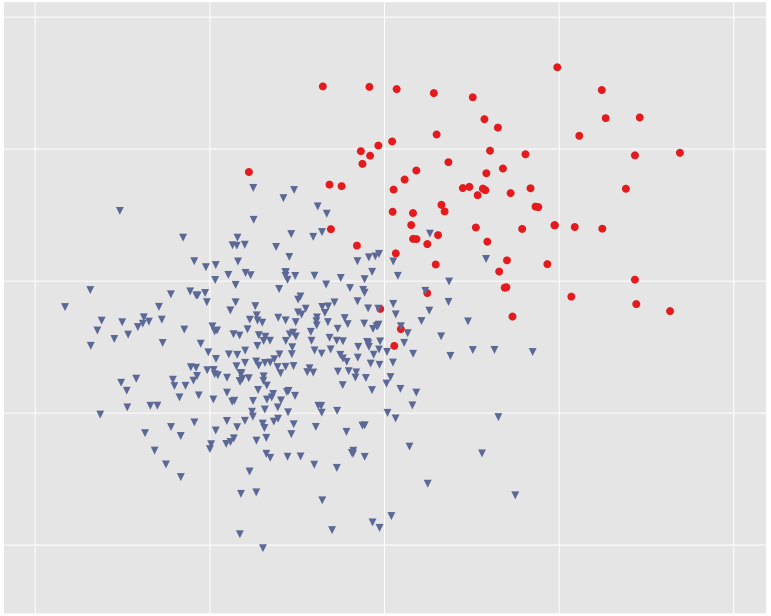
M(C)LE for Logistic Regression

Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of just one test results, X_A and X_B

$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta} \right)$$







Recipe for Estimation

MLE

1. Formulate the likelihood, $p(\mathcal{D} \mid \theta)$
2. Set objective $J(\theta)$ equal to negative log of likelihood

$$\mathrm{J}(\theta) = -\log p(\mathcal{D} \mid \theta)$$

3. Compute derivative of objective, $\partial J / \partial \theta$
4. Find $\hat{\theta}$, either
   a. Set derivate equal to zero and solve for θ
   b. Use (stochastic) gradient descent to step towards better θ

M(C)LE for Multi-class Logistic Regression

Learn to predict if probability of output belonging to class k, Y_k , given input X, $P(Y_k = 1 \mid X, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$

M(C)LE for Multi-class Logistic Regression

Learn to predict if probability of output belonging to class k, Y_k , given input X, $P(Y_k = 1 \mid X, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$

$$\mathcal{L}(\Theta; \mathcal{D}) = \prod_i^N \prod_k^K \left(\frac{e^{\boldsymbol{\theta}_k^T \mathbf{x}^{(i)}}}{\sum_{l=1}^K e^{\boldsymbol{\theta}_l^T \mathbf{x}^{(i)}}} \right)^{\mathbb{I}\left(y_k^{(i)}=1\right)}$$

$\hat{y}_k^{(i)}$

cross ent loss $\hat{y} = sm(X\Theta)$

MLE – Linear Regression

Poll

Implement a function in Python for the pdf of a Gaussian distribution.

Python numpy or math packages are fine, no scipy, etc.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

```
def gaussian(x, mu, sigmaSq):
```

What is `gaussian(3.3, 2.2, 1.1)`?

Poll

Assume that exam scores are drawn independently from the same Gaussian (Normal) distribution.

Given three exam scores {75, 80, 90}, which pair of parameters is a better fit (a higher likelihood)?

A) Mean 80, standard deviation 3

B) Mean 85, standard deviation 7

C) I don't know

$$\mathcal{L}(\mu, \sigma; D) = \prod p(x|\theta)$$

Use a calculator/computer.

Gaussian PDF: $p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$

MLE for Gaussian

Gaussian distribution:

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

$$\mathcal{D} = \{y^{(1)} = 65, y^{(2)} = 95, y^{(3)} = 85\}$$

Formulate the likelihood for three i.i.d. samples, given parameters μ, σ^2?

$$L(\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y^{(i)}-\mu\right)^2}{2\sigma^2}}$$

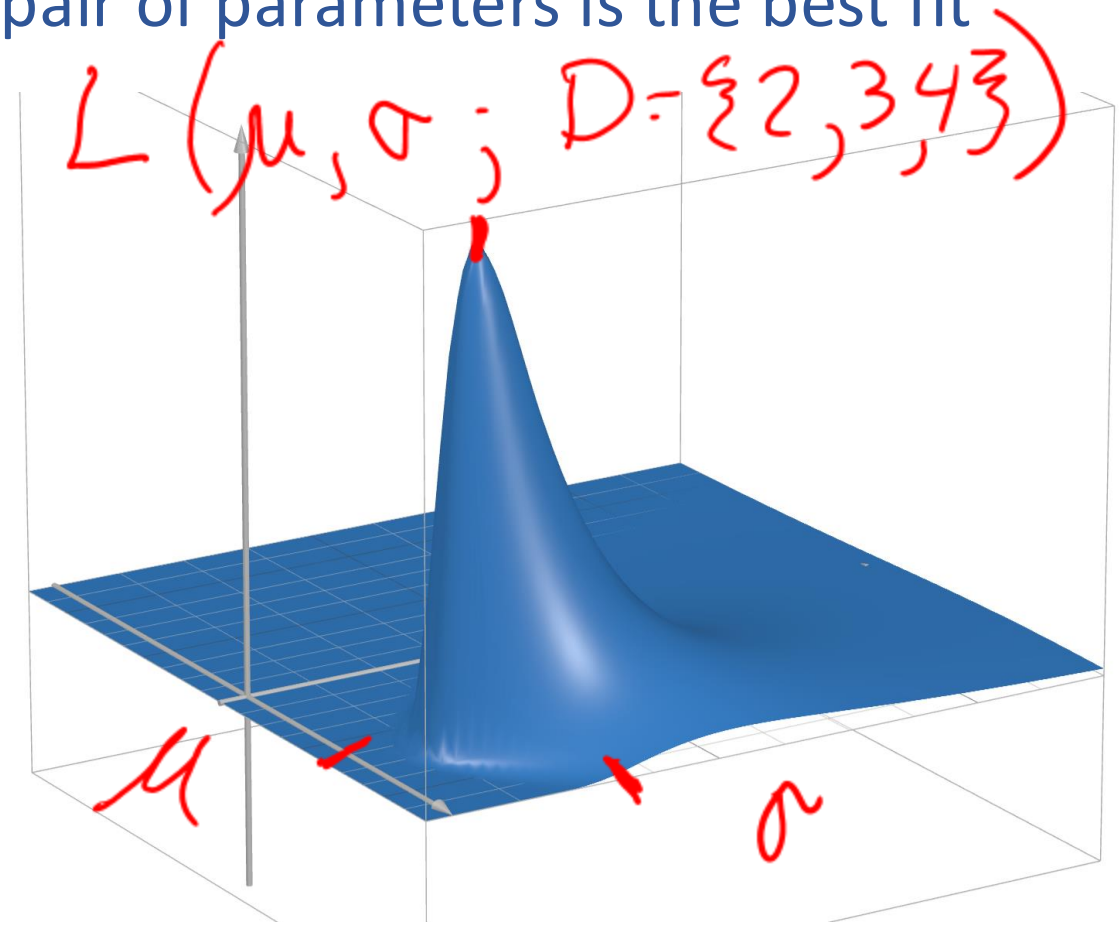
$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} p\left(y^{(i)} \mid \boldsymbol{\theta}\right)$$

MLE for Gaussian

Assume that exam scores are drawn independently from the same Gaussian (Normal) distribution.

Given three exam scores 2, 3, 4, which pair of parameters is the best fit (the highest likelihood)?

$$p(\mathcal{D}|\mu,\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^{(i)}-\mu)^2}{2\sigma^2}}$$



https://www.desmos.com/3d/988327bd26

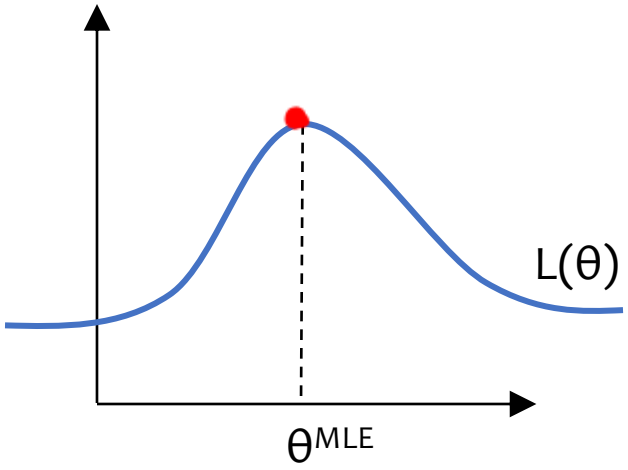
MLE

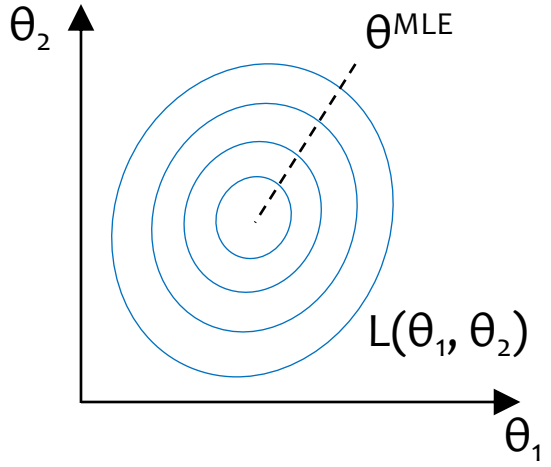
Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Principle of Maximum Likelihood Estimation:
Choose the parameters that maximize the likelihood of the data.

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)

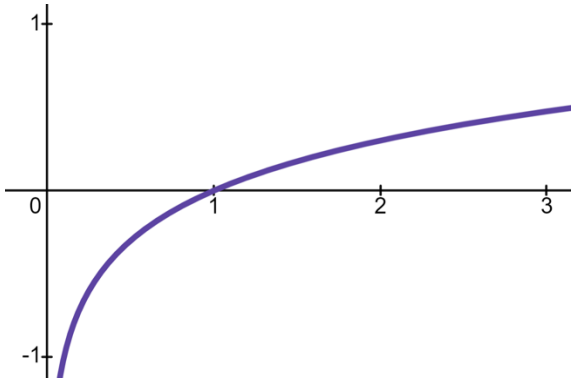




Slide credit: CMU MLD Matt Gormley

Why Log Likelihood?

Log is monotonic:
If $a < b$
then $\log a < \log b$



Updating our objective: Minimize neg. log likelihood

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} \prod_{i=1}^{N} p\left(y^{(i)} \mid \theta\right)$$

$$= \underset{\theta}{\text{argmax}} \sum_{i=1}^{N} \log p\left(y^{(i)} \mid \theta\right)$$

$$= \underset{\theta}{\text{argmin}} -\sum_{i=1}^{N} \log p\left(y^{(i)} \mid \theta\right)$$

Minimize $J(\theta) = -\log \mathcal{L}(\theta; \mathcal{D}) = -\sum_{i=1}^{N} \log p\left(y^{(i)} \mid \theta\right)$

MLE for Gaussian

Gaussian distribution:

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

What is the log likelihood for three i.i.d. samples, given parameters μ, σ^2?

$$\mathcal{D} = \{y^{(1)} = 75, y^{(2)} = 80, y^{(3)} = 90\}$$

$$L(\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y^{(i)}-\mu\right)^2}{2\sigma^2}}$$

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{N} -\log\sqrt{2\pi\sigma^2} - \frac{\left(y^{(i)} - \mu\right)^2}{2\sigma^2}$$

$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} p\left(y^{(i)} \mid \boldsymbol{\theta}\right)$$

$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i}^{N} \log p\left(y^{(i)} \mid \boldsymbol{\theta}\right)$$

Recipe for Estimation

MLE

1. Formulate the likelihood, $p(\mathcal{D} \mid \theta)$
2. Set objective $J(\theta)$ equal to negative log of likelihood

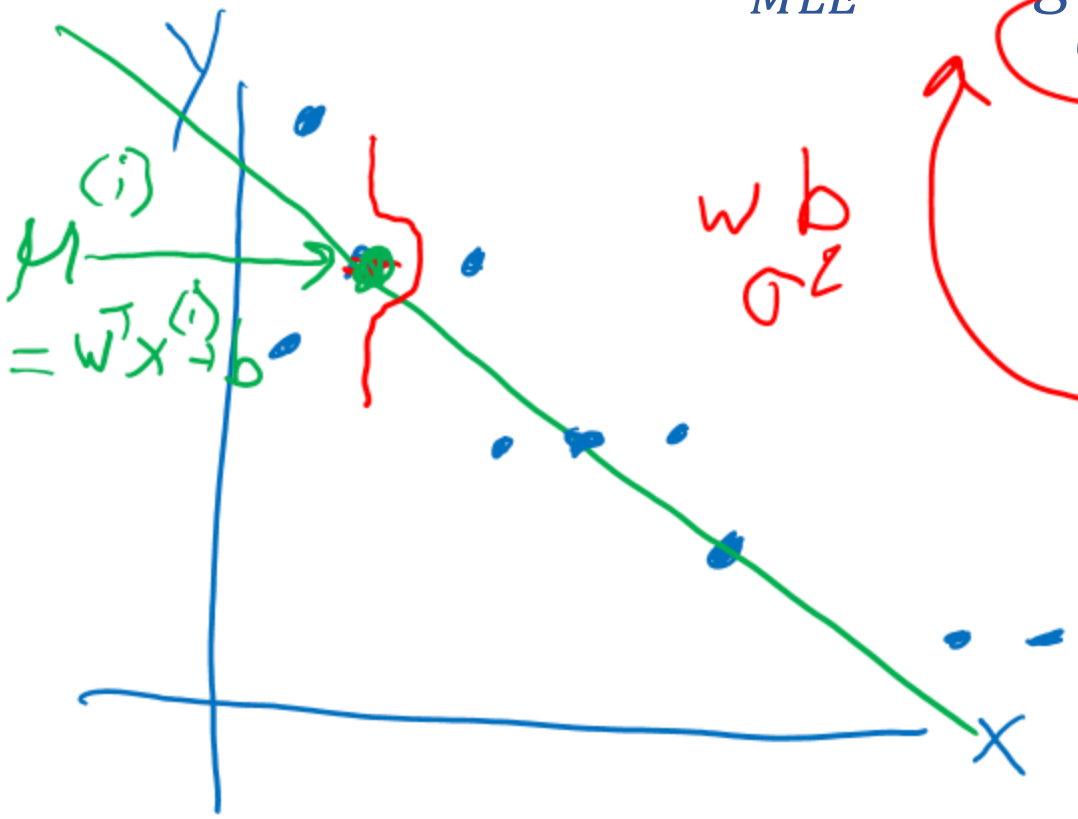
   $$\mathrm{J}(\theta) = -\log p(\mathcal{D} \mid \theta)$$
3. Compute derivative of objective, $\partial J / \partial \theta$
4. Find $\hat{\theta}$, either
   a. Set derivate equal to zero and solve for θ
   b. Use (stochastic) gradient descent to step towards better θ

M(C)LE for Linear Regression

$$f(z;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Probabilistic interpretation of linear regression

$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right)$$



$h(y;x) =$ $y = w^T x + b + \epsilon$

$\epsilon \sim \mathcal{N}(0, \sigma^2)$

$p(y \mid x) \sim \mathcal{N}(w^T x + b, \sigma^2)$

$\mu = w^T x + b$ params

$\sigma^2 \leftarrow$ param

M(C)LE for Linear Regression

Probabilistic interpretation of linear regression

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_i p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right)$$

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(z-\mu)^2}{2\sigma^2}}$$

$$\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(z^{(i)}-\mu\right)^2}{2\sigma^2}}$$

$$\sum_{i=1}^{N} -\log\sqrt{2\pi\sigma^2} - \frac{\left(z^{(i)} - \mu\right)^2}{2\sigma^2}$$

$\sigma^2 = 1$

$\mu^{(i)} = \theta^T x^{(i)}$

$$\ell(\theta; D) = \sum_{i=1}^{N} -\log\sqrt{2\pi\sigma^2} - \frac{\left(y^{(i)} - \mu^{(i)}\right)^2}{2\sigma^2}$$

$$-\ell(\theta; D) = N \log\sqrt{2\pi\sigma^2} + \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y^{(i)} - \theta^T x^{(i)}\right)^2$$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - \theta^T x^{(i)}\right)^2$$