UVA CS 4774: Machine Learning

Lecture 10: Maximum Likelihood Estimation (MLE)

Dr. Yanjun Qi
University of Virginia
Department of Computer Science

Machine Learning in a Nutshell

ML grew out of work in AI

Optimize a performance criterion using example data or past experience,

Aiming to generalize to unseen data

Data
↓
Task
↓
Representation
↓
Score Function
↓
Search/Optimization
↓
Models, Parameters

Hyperparameter, Metrics

Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties, e.g., Independence, conditional independence
- Maximum Likelihood Estimation

Sample space and Events

- O : Sample Space,
  - set of all outcomes
  - If you toss a coin twice O = {HH,HT,TH,TT}
- Event: a subset of O
  - First toss is head = {HH,HT}
- S: event space, a set of events:
  - Contains the empty event and O

From Events to Random Variable

- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - O = all possible students (sample space)
  - What are events (subset of sample space)
    - Grade_A = all students with grade A
    - Grade_B = all students with grade B
    - HardWorking_Yes = … who works hard
  - Very cumbersome

  - Need "functions" that maps from O to an attribute space T.
  - P(H = YES) = P({student ϵ O : H(student) = YES})

If hard to directly estimate from data, most likely we can estimate

- 1. Joint probability
  - Use Chain Rule
- 2. Marginal probability
  - Use the total law of probability
- 3. Conditional probability
  - Use the Bayes Rule

If hard to directly estimate from data, most likely we can estimate

- 1. Joint probability
  - Use Chain Rule
- 2. Marginal probability
  - Use the total law of probability
- 3. Conditional probability
  - Use the Bayes Rule

$$P(A, B) = P(B)\, P(A|B)$$

$$P(B) = P(B, A) + P(B, \sim A)$$

$$= P(B, A \cup \sim A)$$

$$P(A|B)$$

$$P(B|A) = \frac{P(A, B)}{P(A)} = \frac{P(A|B)\, P(B)}{P(A)}$$

Simplify Notation: To Calculate Conditional Probability

- Bayes Rule

$$P(x \mid y) = \frac{P(x)P(y \mid x)}{P(y)}$$

- You can condition on more variables

$$P(x \mid y, z) = \frac{P(x \mid z)P(y \mid x, z)}{P(y \mid z)}$$

Examples

Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the **set** *{r,r,r,b}. What is the probability of drawing 2 red balls in the first 2 tries?*

$$P(B_1 = r, B_2 = r) =$$

What is the probability that the 2nd ball drawn from the **set** *{r,r,r,b} will be red?*

$$P(B_2 = r)$$

Adapt from Prof. Nando de Freitas's review slides

Today : MLE

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties, e.g., Independence, conditional independence
- Maximum Likelihood Estimation

Roadmap

- ❑ Basic MLE
- ❑ MLE for Discrete RV
- ❑ MLE for Continuous RV (Gaussian)
- ❑ MLE connects to Normal Equation of LR
- ❑ More about Mean and Variance

Review: Maximum Likelihood Estimation

A general Statement

Consider a sample set $T=(Z_1...Z_n)$ which is drawn from a probability distribution P(Z|\theta) where \theta are parameters.

If the Zs are independent with probability density function $P(Z_i|\theta)$, the joint probability of the whole set is

$$P(Z_1 ... Z_n|\theta) = \prod_{i=1}^{n} P(Z_i|\theta)$$

this may be maximised with respect to \theta to give the maximum likelihood estimates.

The idea is to

✓ assume a particular model with unknown parameters, θ

The idea is to

- ✓ assume a particular model with unknown parameters, θ
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters. $P(Z_i|\theta)$

The idea is to

- ✓ assume a particular model with unknown parameters, θ
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters. $P(Z_i|\theta)$
- ✓ We have observed a set of outcomes in the real world.

The idea is to

- ✓ assume a particular model with unknown parameters, θ
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters. $P(Z_i|\theta)$
- ✓ We have observed a set of outcomes in the real world. $z_1, z_2, \ldots, z_n$
- ✓ It is then possible to choose a set of parameters which are most likely to have produced the observed results.

The idea is to

- ✓ assume a particular model with unknown parameters, θ
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters. $P(Z_i|\theta)$
- ✓ We have observed a set of outcomes in the real world.
- ✓ It is then possible to choose a set of parameters which are most likely to have produced the observed results.

$$\hat{\theta} = \underset{\theta}{argmax} \; P(Z_1 \ldots Z_n|\theta) = \prod_{i=1}^{n} P(Z_i|\theta)$$

This is maximum likelihood. In most cases it is both consistent and efficient.

$$log(L(\theta)) = \sum_{i=1}^{n} log(P(Z_i|\theta))$$

It is often convenient to work with the Log of the likelihood function.

The idea is to

- ✓ assume a particular model with unknown parameters, θ
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters. $P(Z_i|\theta)$
- ✓ We have observed a set of outcomes in the real world.
- ✓ It is then possible to choose a set of parameters which are most likely to have produced the observed results.

$$\hat{\theta} = \underset{\theta}{argmax} \; P(Z_1 \ldots Z_n|\theta)$$

Likelihood

This is maximum likelihood.

In most cases this scorer is both consistent and efficient.

$$log(L(\theta)) = \sum_{i=1}^{n} log(P(Z_i|\theta))$$

Log-Likelihood

It is often convenient to work with the Log of the likelihood function.

Roadmap

- ❑ Basic MLE
- ❑ MLE for Discrete RV
- ❑ MLE for Continuous RV (Gaussian)
- ❑ MLE connects to Normal Equation of LR
- ❑ More about Mean and Variance

Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g. Z as the total number of heads you get if you flip 100 coins

- Z is a RV with arity *k* if it can take on exactly one value out of *a set size k*
  - E.g. the possible values that Z can take on are 0, 1, 2,..., 100

Review: Bernoulli Distribution e.g. Coin Flips

- You flip a coin
  - Z: {Who is Up: Head or Tail} is a discrete RV
  - Head with probability *p*
  - Binary random variable
  - Bernoulli trial with success probability *p*

$\{H, T\}$

Review: Bernoulli Distribution e.g. Coin Flips

- You flip *n* coins
  - Head with probability *p* (*UNKNOWN, Need to estimate from data*)
  - Number of heads X out of n trial
  - Each Trial following Bernoulli distribution with parameters *p*

$$\hat{\theta} = \underset{\theta}{argmax} \; P\left(Z_1 \ldots Z_n | \theta\right)$$

Review: Defining Likelihood for basic Bernoulli

Given: $\{Z_1, Z_2, \ldots, Z_n\}$

$\Downarrow$

$\{H, H, T, \ldots H\}_n$

$\Downarrow$ reformulate

$\{1, 1, 0, \ldots, 1\}_n$

$p(Z_i | \underline{\underline{\theta}}) = p^{Z_i}(1-p)^{1-Z_i}$ (Here $Z_i \in \{0, 1\}$)

$$p(Z_i) = \begin{cases} p, & \text{if } Z_i = H/1 \\ 1-p, & \text{if } Z_i = T/0 \end{cases} \Rightarrow \underset{p}{\text{argmax}} \prod_{i=1}^{n} p^{Z_i}(1-p)^{1-Z_i}$$

$\theta = \{p\}$

$= \{p(\text{Head})\}$

Defining Likelihood

Observing binary samples z_i

PMF:

$$\Pr(z_i|p) = p^{z_i}(1-p)^{1-z_i}$$

Observed data ➔ x heads-up from n trials

LIKELIHOOD:

$$L(p) = \prod_{i=1}^{n} p^{z_i}(1-p)^{1-z_i}$$

function of p=Pr(head)

$$\log(L(p)) = \log\left[\prod_{i=1}^{n} p^{z_i}(1-p)^{1-z_i}\right]$$

$$= \sum_{i=1}^{n}(z_i \log p + (1-z_i)\log(1-p))$$

Deriving the Maximum Likelihood Estimate for Bernoulli

Minimize the negative log-likelihood

$$\underset{p}{\operatorname{argmin}} \left\{ -l(p) \right\} = -\log(L(p)) = -\log\left[p^x (1-p)^{n-x} \right]$$

$$= -\log(p^x) - \log((1-p)^{n-x})$$

$$= -x\log(p) - (n-x)\log(1-p)$$

Deriving the Maximum Likelihood Estimate for Bernoulli

$$\arg\max_p \{-l(p)\} = \arg\min_p \{-x\log(p)-(n-x)\log(1-p)\}$$

$$\frac{dl(p)}{dp} = -\frac{x}{p} - \frac{-(n-x)}{1-p} = 0$$

$$0 = -\frac{x}{p} + \frac{n-x}{1-p}$$

$$0 = \frac{-x(1-p)+p(n-x)}{p(1-p)}$$

$$0 = -x + px + pn - px$$

$$0 = -x + pn$$

Minimize the negative log-likelihood

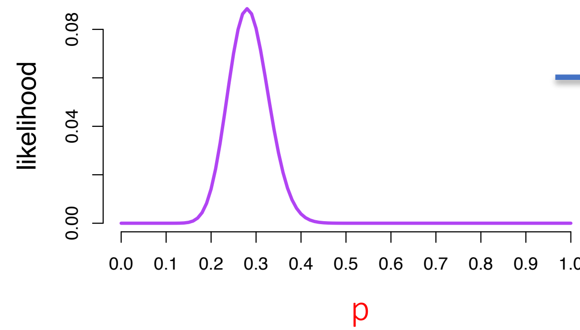
➔ MLE parameter estimation

$$\hat{p} = \frac{x}{n}$$

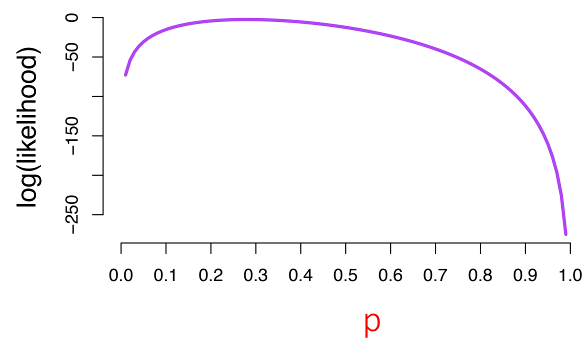
i.e. Relative frequency of a binary event

Deriving the Maximum Likelihood Estimate for Bernoulli

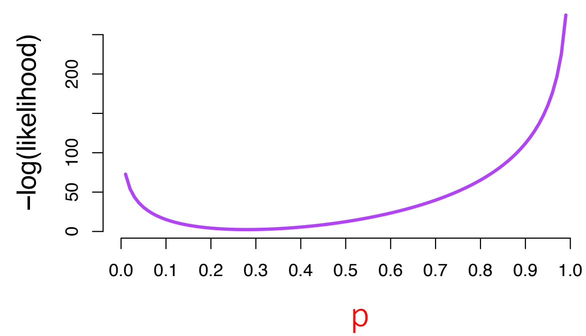
$$\underbrace{L(p)}_{\text{maximize}} = p^x(1-p)^{n-x}$$



$$\underbrace{\log(L(p))}_{\text{maximize}} = \log\left[p^x(1-p)^{n-x}\right]$$



$$\underbrace{-l(p)}_{\text{Minimize the negative log-likelihood}} = -\log\left[p^x(1-p)^{n-x}\right]$$



EXTRA

Roadmap - All the rest are EXTRA

- ❑ Basic MLE
- ❑ MLE for Discrete RV
- ❑ MLE for Continuous RV (Gaussian)
- ❑ MLE connects to Normal Equation of LR
- ❑ More about Mean and Variance

Review: Continuous Random Variables

- Probability density function (pdf) instead of probability mass function (pmf)
  - For discrete RV: Probability mass function (pmf): $P(X = x_i)$

- A pdf (prob. Density func.) is any function $f(x)$ that describes the probability density in terms of the input variable x.

Review: Probability of Continuous RV

- Properties of pdf
  - $$f(x) \geq 0, \forall x$$
  - $$\int_{-\infty}^{+\infty} f(x) = 1$$

$$\longrightarrow \sum_{i=1}^{k_i} P(X = x_i) = 1$$

- Actual probability can be obtained by taking the integral of pdf
  - E.g. the probability of X being between 5 and 6 is

$$P(5 \leq X \leq 6) = \int_{5}^{6} f(x)dx$$

Review: Mean and Variance of RV

- Mean (Expectation):
  $$\mu = E(\mathrm{X})$$
  - Discrete RVs:
  $$E(\mathrm{X}) = \sum_{v_i} v_i \mathrm{P}(\mathrm{X} = v_i)$$
  $$E(g(X)) = \sum_{v_i} g(v_i) P(X = v_i)$$
  - Continuous RVs:
  $$E(\mathrm{X}) = \int_{-\infty}^{+\infty} x f(x) dx$$
  $$E(g(X)) = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

Adapt From Carols' prob tutorial

Review: Mean and Variance of RV

- Variance: $$Var(X) = E((X-\mu)^2)$$

$\sigma_X = \sqrt{V(X)}$

  - Discrete RVs:

$$V(X) = \sum_{v_i} (v_i - \mu)^2 P(X = v_i)$$

  - Continuous RVs:

$$V(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx$$

Correlation

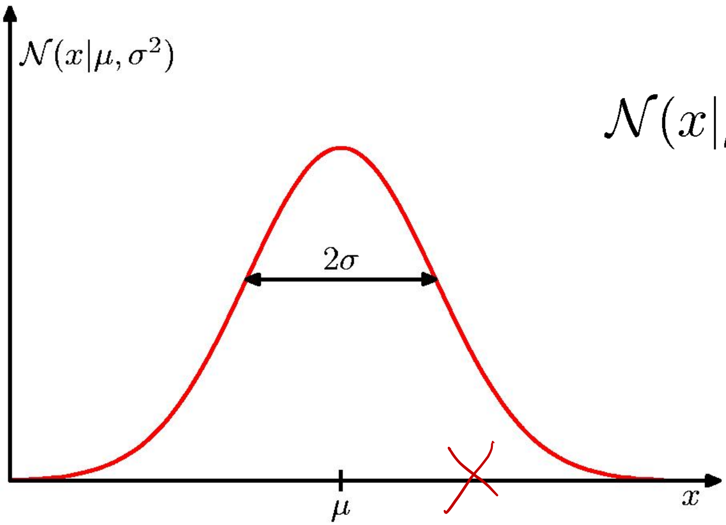
$\rho_{X,Y} = Corr(X,Y) = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y}$

- Covariance:

$$Cov(X,Y) = E((X-\mu_x)(Y-\mu_y)) = E(XY) - \mu_x \mu_y$$

Adapt From Carols' prob tutorial

Single-Variate Gaussian Distribution

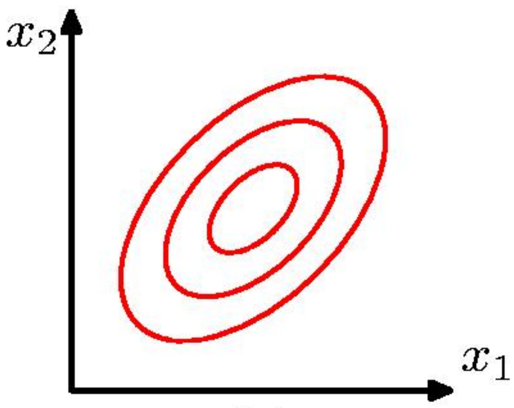


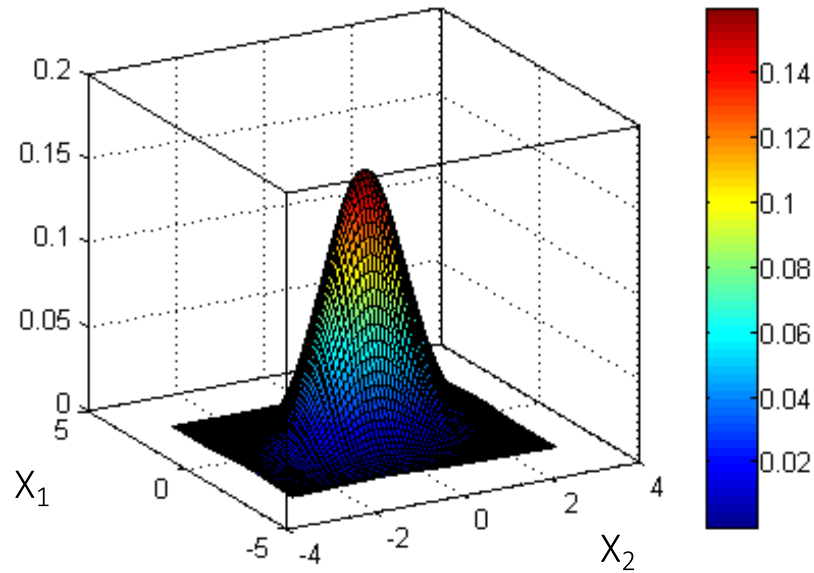
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$X \sim N\left(\mu, \sigma^2\right)$$

Courtesy: http://research.microsoft.com/~cmbishop/PRML/index.htm

Bi-Variate Gaussian Distribution





Bivariate normal PDF:

- Mean of normal PDF is at peak value. Contours of equal PDF form ellipses.

- The covariance matrix captures linear dependencies among the variables

Courtesy: http://research.microsoft.com/~cmbishop/PRML/index.htm

Multivariate Normal (Gaussian) PDFs

The only widely used continuous joint PDF is the multivariate normal (or Gaussian):

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Mean

Covariance Matrix

Where |*| represents determinant

- Mean of normal PDF is at peak value. Contours of equal PDF form ellipses.

- The covariance matrix captures linear dependencies among the variables

Example: the Bivariate Normal distribution

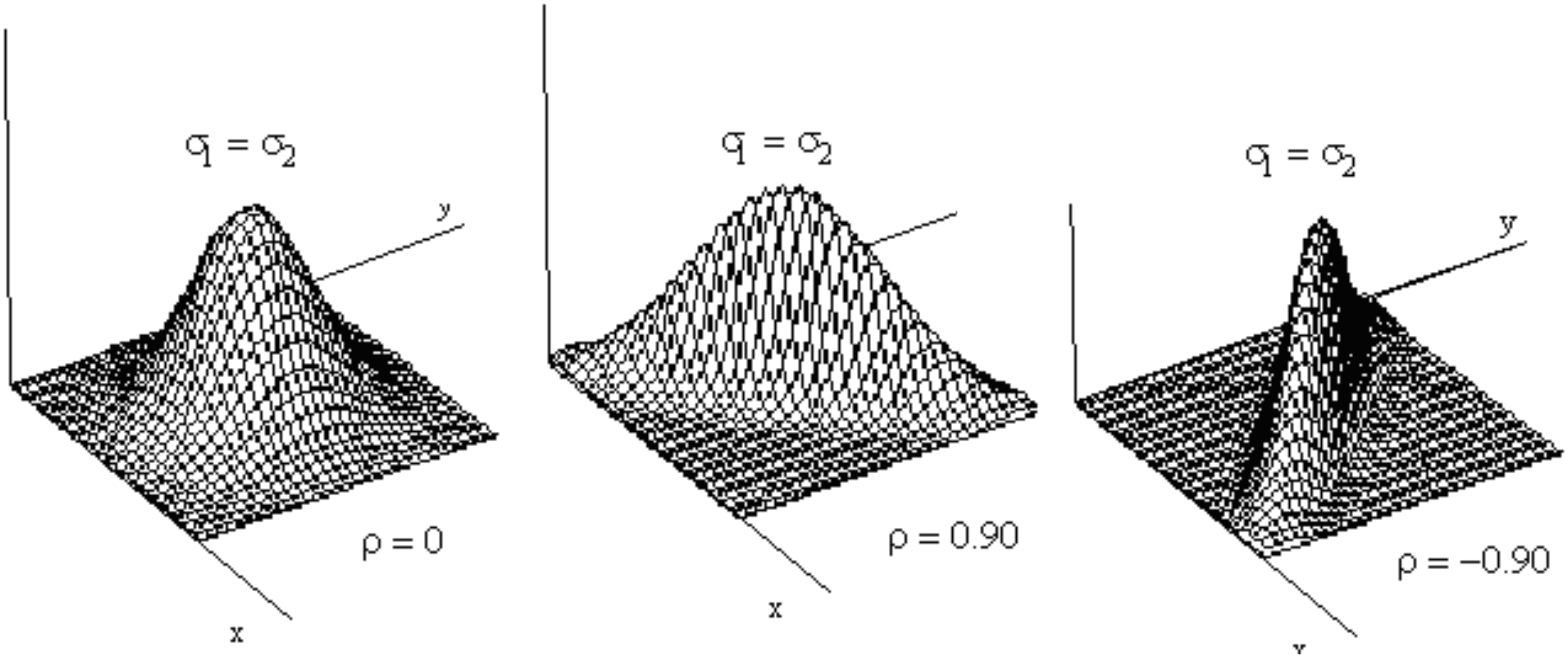
$$f\left(x_1,x_2\right)=\frac{1}{\left(2\pi\right)\left|\Sigma\right|^{1/2}}\mathrm{e}^{-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T\Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)}$$

with $\vec{\mu}=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}_{2\times 1}$ and

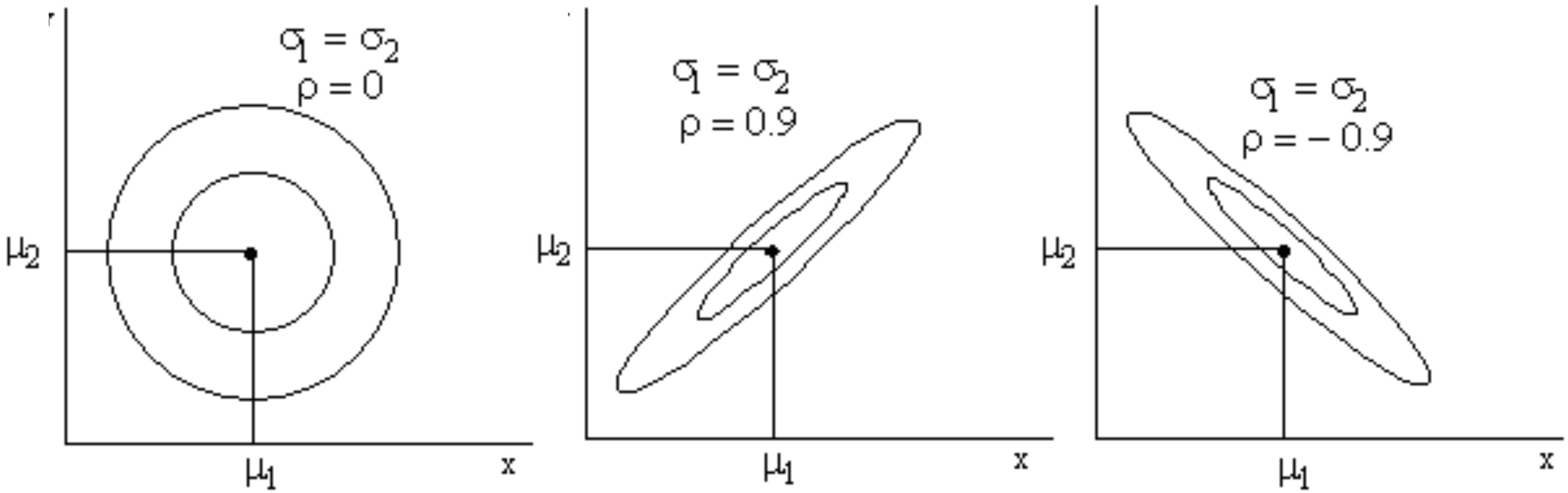
$$\underset{2\times 2}{\Sigma}=\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{bmatrix}=\begin{bmatrix}\overset{V(X_1)}{\sigma_1^2} & \overset{Cov(X_1,X_2)}{\underline{\rho\sigma_1\sigma_2}}\\ \underline{\rho\sigma_1\sigma_2} & \underset{V(X_2)}{\sigma_2^2}\end{bmatrix}_{2\times 2}$$

$$\left|\Sigma\right|=\sigma_{11}\sigma_{22}-\sigma_{12}^2=\sigma_1^2\sigma_2^2\left(1-\rho^2\right)$$

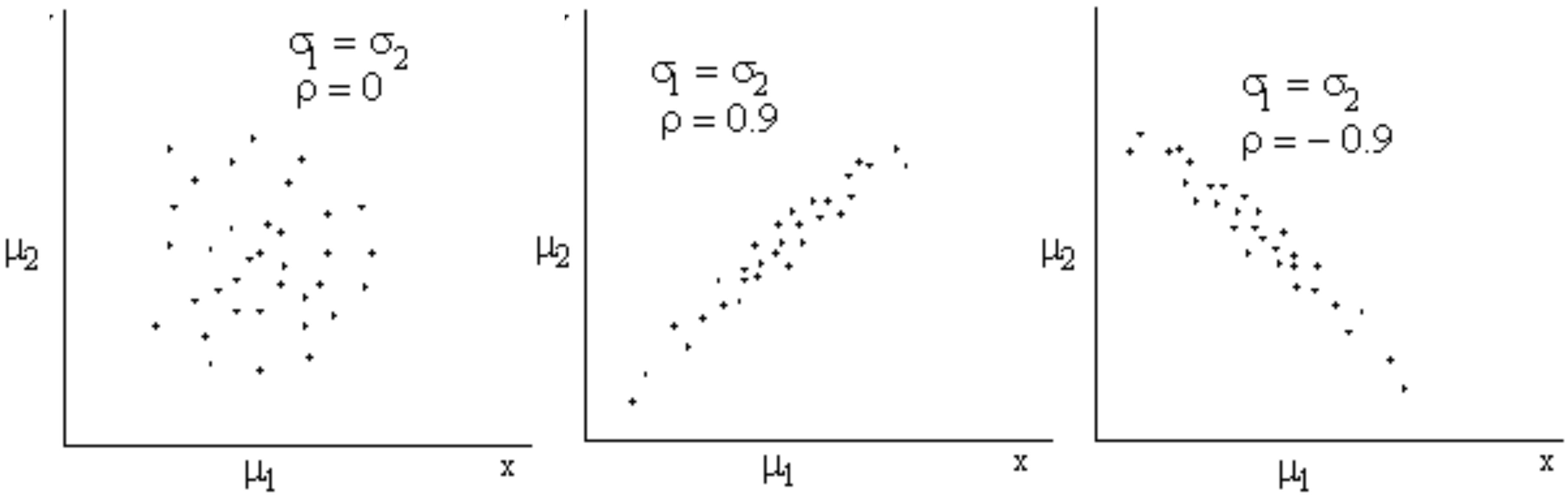
Surface Plots of the bivariate Normal distribution



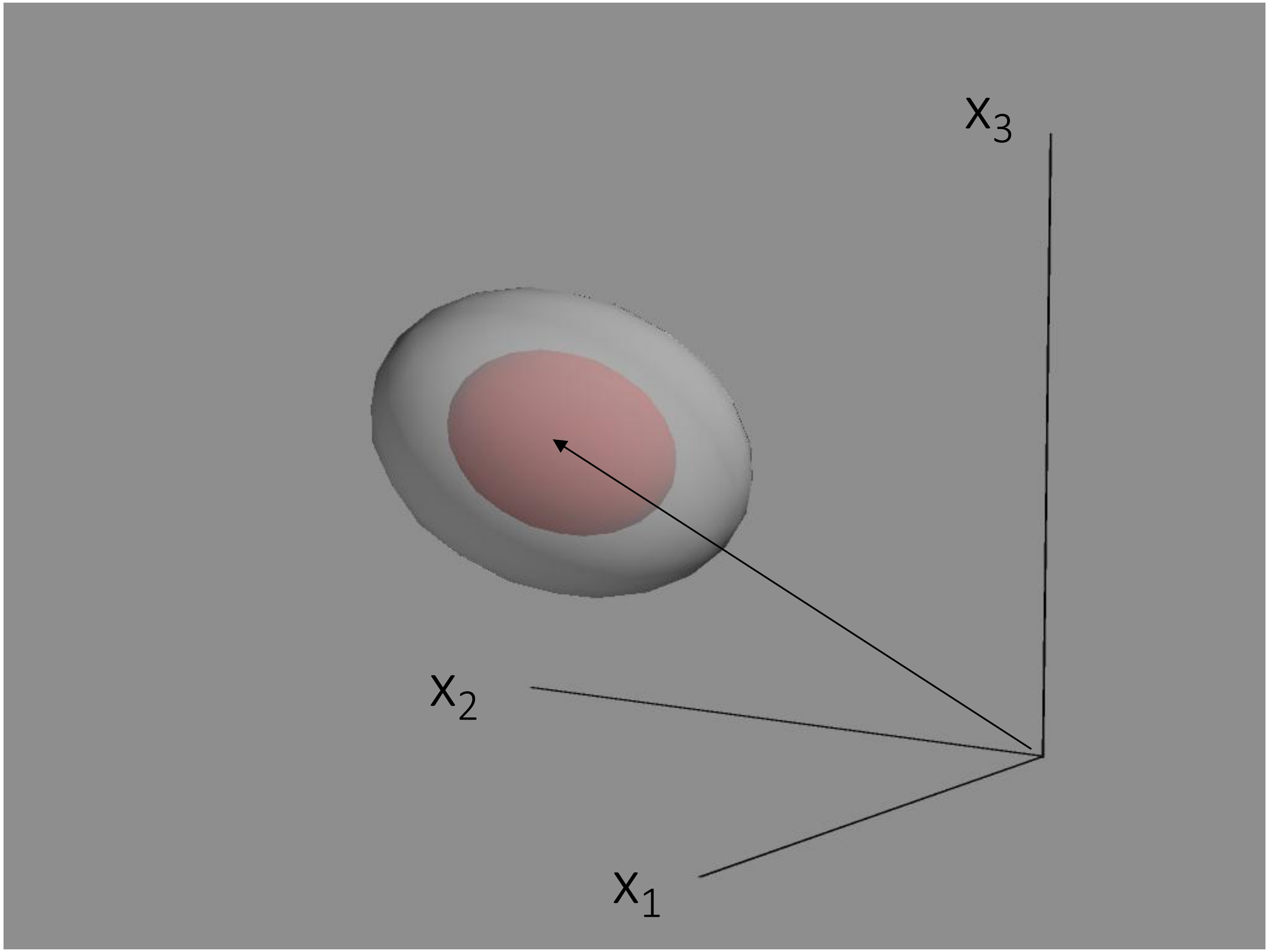
Contour Plots of the bivariate Normal distribution



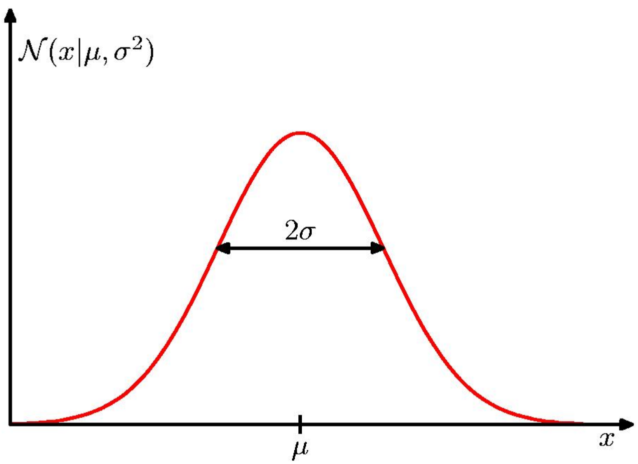
Scatter Plots of data from the bivariate Normal distribution



Trivariate Normal distribution



How to Estimate 1D Gaussian: MLE



- In the 1D Gaussian case, we simply set the mean and the variance to the sample mean and the sample variance:

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{\mu}\right)^2$$

How to Estimate p-D Gaussian: MLE

$$< X_1, X_2 \cdots, X_p > \sim N\left(\vec{\mu}, \Sigma\right)$$

$$\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}_{p \times 1}$$

$$\mu_i = \frac{1}{n} \sum_{j=1}^{N} \underbrace{X_j^{(i)}}_{j\text{th sample}}$$

i-th feature $\in \{1, 2, \cdots, p\}$

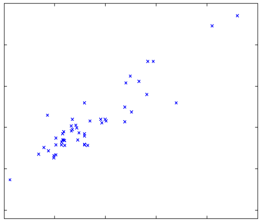
jth sample $\in \{1, 2, \cdots, N\}$

$$\underset{p \times p}{\Sigma} = \begin{bmatrix} Var(X_1) & & \cdots & Cov(X_i, X_j) \\ & \ddots & & \\ & & & Var(X_p) \end{bmatrix}$$

Today

- ❑ Basic MLE
- ❑ MLE for Discrete RV
- ❑ MLE for Continuous RV (Gaussian)
- ❑ MLE connects to Normal Equation of LR
- ❑ More about Mean and Variance

DETOUR: Probabilistic Interpretation of Linear Regression

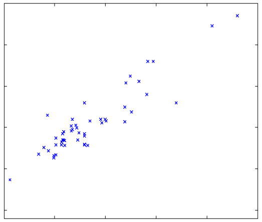


- Let us assume that the target variable and the inputs are related by the equation:

$$y_i = \theta^T \mathbf{x}_i + \varepsilon_i$$

where ε is an error term of unmodeled effects or random noise

DETOUR: Probabilistic Interpretation of Linear Regression



- Let us assume that the target variable and the inputs are related by the equation:

$$y_i = \theta^T \mathbf{x}_i + \varepsilon_i$$

RV $\varepsilon \sim N(0, \sigma^2)$

where ε is an error term of unmodeled effects or random noise

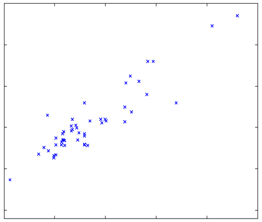
2

- Now assume that ε follows a Gaussian N(0,σ), then we have:

$$p(y_i \mid x_i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \theta^T \mathbf{x}_i)^2}{2\sigma^2} \right)$$

RV $y|x;\theta \sim N(\theta^T x, \sigma)$

DETOUR: Probabilistic Interpretation of Linear Regression



- By IID (independent and identically distributed) assumption, we have data likelihood

$$L(\theta)=\prod_{i=1}^{n}p(y_i\,|\,x_i;\theta)=\left(\frac{1}{\sqrt{2\pi\sigma}}\right)^{n}\exp\left(-\frac{\sum_{i=1}^{n}(y_i-\theta^T\mathbf{x}_i)^2}{2\sigma^2}\right)$$

$$l(\theta)=\log(L(\theta))=n\log\frac{1}{\sqrt{2\pi\sigma}}-\frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta^T\mathbf{x}_i)^2$$

$$L(\theta) = \prod_{i=1}^{n} p(y_i | x_i; \theta) = \left(\frac{1}{\sqrt{2\pi\sigma}} \right)^n \exp\left(- \frac{\sum_{i=1}^{n} (y_i - \theta^T \mathbf{x}_i)^2}{2\sigma^2} \right)$$

We can learn \theta by maximizing the probability / likelihood of generating the observed samples:

$$p\{(\vec{x}_1, y_1) \wedge (\vec{x}_2, y_2) \wedge \cdots (\vec{x}_N, y_N)\}$$

IID

$$= \prod_{i=1}^{N} p(y_i, \vec{x}_i) = \prod_{i=1}^{N} p(y_i | \vec{x}_i; \theta) p(\vec{x}_i)$$

$$\theta^* = \arg\max_{\theta} \prod_{i=1}^{N} p(y_i | \vec{x}_i; \theta)$$

Thus under independence Gaussian residual assumption, residual square error is equivalent to MLE of θ !

$y|x;\theta \sim N(\theta^T x, \sigma)$

$\Downarrow$

Two unknown parameters : $\{\theta, \sigma\}$

$$l(\theta) = \log(L(\theta)) = n \log \frac{1}{\sqrt{2\pi\sigma}} - \frac{1}{\sigma^2} \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta^T \mathbf{x}_i)^2$$

$\Downarrow$

argmax $l(\theta) \Rightarrow$ argmin $J(\theta)$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i^T \theta - y_i)^2$$

$\boxed{y_i \sim N(exp(wx_i), 1)}$

(b) (6 points) (no explanation required) Suppose you decide to do a maximum likelihood estimation of w. You do the math and figure out that you need w to satisfy one of the following equations. Which one?

A. $\Sigma_i x_i exp(wx_i) = \Sigma_i x_i y_i exp(wx_i)$

B. $\Sigma_i x_i exp(2wx_i) = \Sigma_i x_i y_i exp(wx_i)$

C. $\Sigma_i x_i^2 exp(wx_i) = \Sigma_i x_i y_i exp(wx_i)$

D. $\Sigma_i x_i^2 exp(wx_i) = \Sigma_i x_i y_i exp(wx_i/2)$

E. $\Sigma_i exp(wx_i) = \Sigma_i y_i exp(wx_i)$

$y_i \sim N(exp(wx_i), 1)$

Answer: B (this is an extra credit question.)

$$L(\theta)$$
$$\downarrow$$
$$LL(\theta)$$
$$\downarrow$$
$$\frac{\partial LL(\theta)}{\partial \theta} = 0 \Rightarrow (B)$$

Today

- ❑ Basic MLE
- ❑ MLE for Discrete RV
- ❑ MLE for Continuous RV (Gaussian)
- ❑ MLE connects to Normal Equation of LR
- ❑ Extra: about Mean and Variance

Mean and Variance

- Correlation:

$$\rho(X,Y) = Cov(X,Y)/\sigma_x\sigma_y$$

$$-1 \leq \rho(X,Y) \leq 1$$

Properties

- Mean $$E(\mathrm{X}+\mathrm{Y})=E(\mathrm{X})+E(\mathrm{Y})$$ $$E(a\mathrm{X})=aE(\mathrm{X})$$
  - If X and Y are independent, $$E(\mathrm{XY})=E(\mathrm{X})\cdot E(\mathrm{Y})$$
- Variance $$V(a\mathrm{X}+b)=a^2V(\mathrm{X})$$
  - If X and Y are independent, $$V(\mathrm{X}+\mathrm{Y})=V(\mathrm{X})+V(\mathrm{Y})$$

Some more properties

- The conditional expectation of Y given X when the value of X = x is:

$$E(Y \mid X = x) = \int y * p(y \mid x) dy$$

- The Law of Total Expectation or Law of Iterated Expectation:

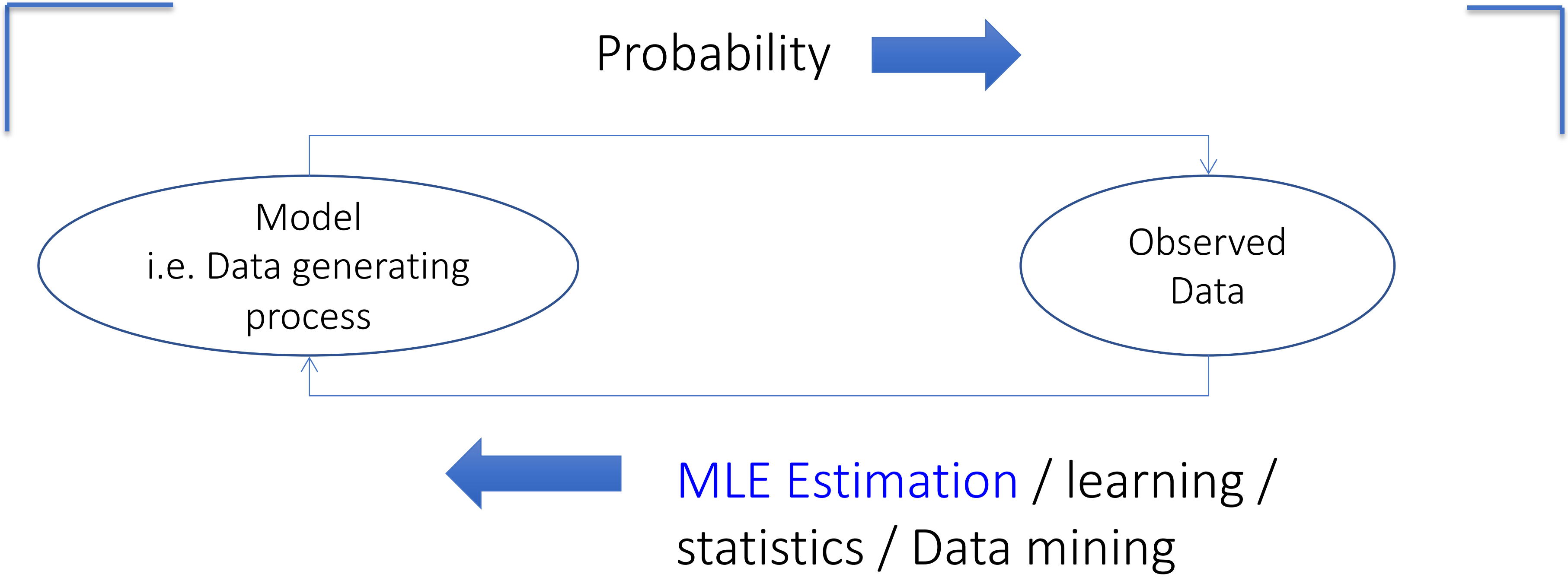
$$E(Y) = E[E(Y \mid X)] = \int E(Y \mid X = x) p_X(x) dx$$

Some more properties

- The law of Total Variance:

$$Var(Y) = Var\left[E(Y \mid X)\right] + E\left[Var(Y \mid X)\right]$$

The Big Picture



e.g. Coin Flips cont.

$\{H, T\}$

- You flip a coin
  - Z: {Who is Up: Head or Tail} is a discrete RV
  - Head with probability *p*
  - Binary random variable
  - Bernoulli trial with success probability *p*
- You flip *a* coin for *k* times
  - How many heads would you expect
  - Number of heads Z is also a discrete random variable
  - Binomial distribution with parameters *k* and *p*

References

- Prof. Andrew Moore's review tutorial
- Prof. Nando de Freitas's review slides
- Prof. Carlos Guestrin recitation slides