Machine Learning

Lecture 13: Maximum Likelihood Estimation

Dr. Beilun Wang

Southeast University

School of Computer Science and Engineering

Last: 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

Sample space and Events

- O: Sample Space,
  - result of an experiment / 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 (RV)

- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - Ω = 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 Ω to an attribute space T.
  - P(H = YES) = P({student ϵ Ω : H(student) = YES})

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

- Joint probability
  - Use Chain Rule

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

- Marginal probability
  - Use the total law of probability

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

- Conditional probability
  - Use the Bayes Rule

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

Today

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

Maximum Likelihood Estimation

- A general Statement
  - Consider a sample set $T = (X_1, \ldots, X_{\mathrm{n}})$ which is drawn from a probability distribution $p(X|\theta)$ where θ are parameters.
  - If the Xs are independent with probability density function $p(X_i|\theta)$, the joint probability of the whole set is

$$p(X_1, \ldots, X_{\mathrm{n}}|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$$

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

Maximum Likelihood Estimation

- 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.
- 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.

$$\hat{\theta} = argmax_{\theta} p(X_1, \dots, X_n | \theta)$$

Likelihood

- This is maximum likelihood.

$$\log\big(L(\theta)\big) = \sum_{i=1}^{n} \log(p(X_i|\theta))$$

Log-Likelihood

- It's often both consistent and efficient.
- It provides a standard to compare other estimation techniques.

Today

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

Discrete Random Variables

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

- X is a RV with arity *k* if it can take on exactly one value out of $\{x_1, \dots, x_k\}$
  - E.g. the possible values that X can take on are 0, 1, 2,…, 100

e.g. Coin Flips cont.

- You flip a coin
  - 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 X is a discrete random variable
  - Binomial distribution with parameters *k* and *p*

Review: Bernoulli Distribution e.g. Coin Flips

- You flip *n* coins
  - How many heads would you expect
  - Head with probability *p*
  - Number of heads X out of n trial
  - Each trial following Bernoulli distribution with parameters *p*

$$N\ trials,$$
$$e.g. \{H, H, T, H, H, T, H, T, \dots, H\}$$
$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, \dots, x_n$$

Calculating Likelihood

$$Given: \ \{x_1, x_2, \dots, x_n, \}$$

$$\Downarrow$$

$$\{H, H, T, \dots, H\}$$

$$\Downarrow \text{ reformulate}$$

$$\{1,1,0, \dots, 1\}$$

$$p(x_i|\theta) = p^{x_i}(1-p)^{1-x_i},\ x_i \epsilon \{0,1\}$$

Defining Likelihood for Bernoulli

- Likelihood = p(data|parameter)
- e.g., for n independent tosses of coins, with unknown parameter p

PMF: $f(x_i \mid p) = p^{x_i}(1-p)^{1-x_i}$

$$x = \sum_{i=1}^{n} x_i$$

Likelihood:

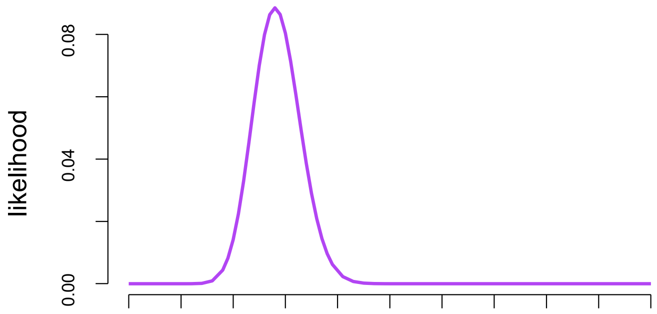
$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^x (1-p)^{n-x}$$

Observed data → x heads-up from n trials

Deriving the Maximum Likelihood Estimate for Bernoulli

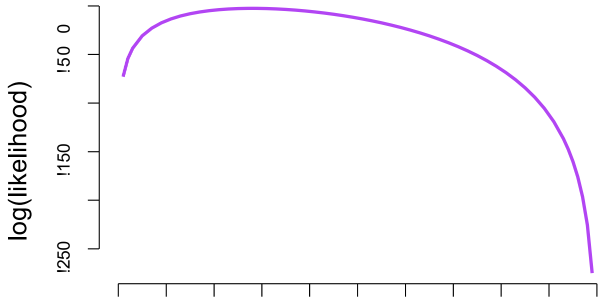
maximize

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



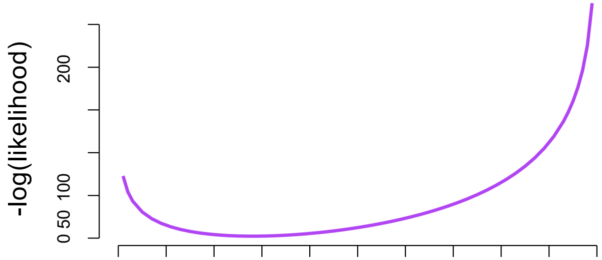
maximize

$$\log\big(L(p)\big) = \log[p^x(1-p)^{n-x}]$$



minimize the negative log-likelihood

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



Deriving the Maximum Likelihood Estimate for Bernoulli

minimize the negative log-likelihood

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

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

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

Deriving the Maximum Likelihood Estimate for Bernoulli

$$\underset{p}{\operatorname{argmin}}\{-l(p)\} = \underset{p}{\operatorname{argmin}}\{-x\log(p) - (n-x)\log(1-p)\}$$

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

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

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

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

→ MLE parameter estimation

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

i.e. Relative frequency of a binary event

Today

- Basic MLE
- MLE for Discrete RV
- MLE for Continuous RV (Gaussian)
- MLE connects to Normal Equation of LR
- Extra: Properties 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

$$f(x) \geq 0, \forall x$$

$$\int_{-\infty}^{+\infty} f(x) = 1$$

- 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):
  - Discrete RVs:

$$\mu = E(X)$$

$$E(X) = \sum_{v_i} v_i P(X = v_i)$$

$$E(g(X)) = \sum_{v_i} g(v_i) P(X = v_i)$$

  - Continuous RVs:

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

$$E(g(X)) = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

Review: Mean and Variance of RV

- Variance:

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

  - 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:

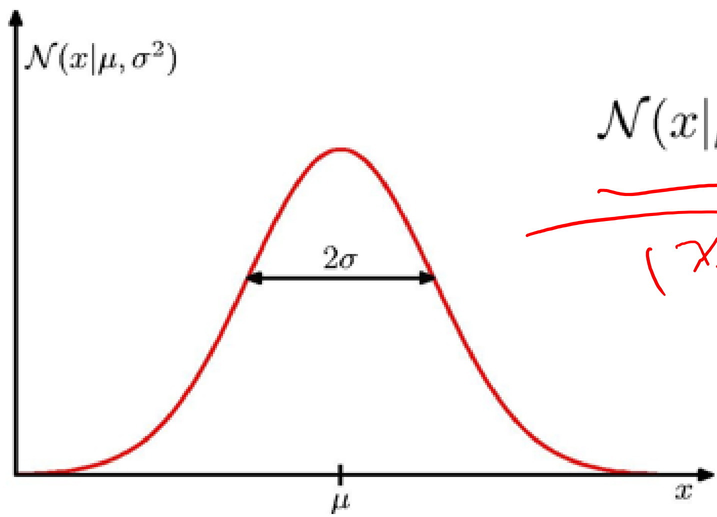
$$\rho_{X,Y} = Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- Covariance:

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

Single-Variate Gaussian Distribution

$\nabla \log \mathcal{L}(\theta) = 0$



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

$(x_1, \cdots, x_n)$ $\quad f(x_1)\cdots f(x_n)$

$\mathcal{L}(\mu,\sigma) = \prod f(x_i)$

$\nabla \log \mathcal{L}(\theta) = 0$, $\theta = (\mu,\sigma)$

$\log \prod f(x_i) = 0 \Rightarrow \nabla \sum \log f(x_i) = 0$

$\nabla \sum \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$

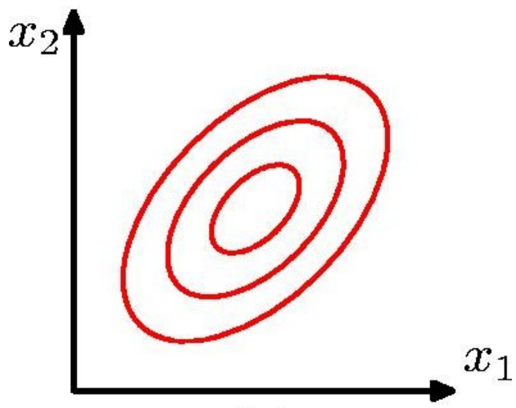
$$X \sim N(\mu,\sigma^2)$$

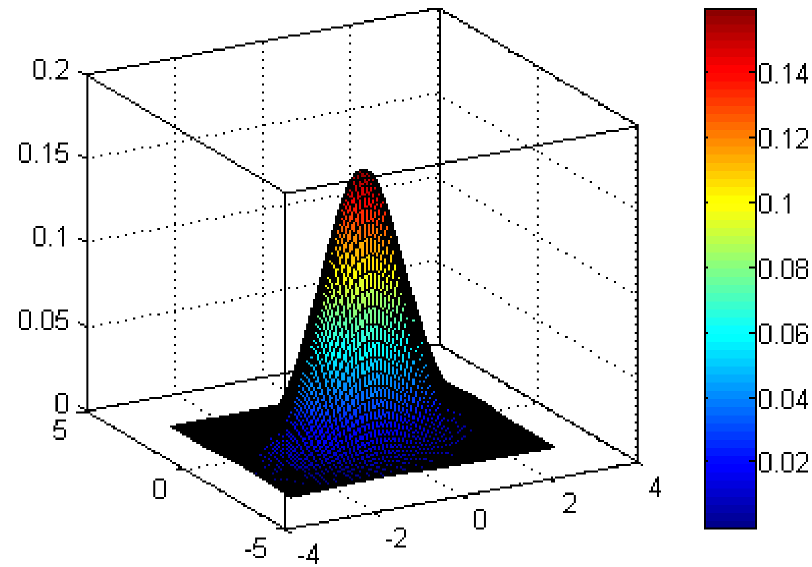
$\nabla_\sigma \quad -\frac{n \cdot 4\pi\sigma}{2 \cdot 2\pi\sigma^2} + \frac{\sum(x_i-\mu)^2}{2} \cdot \frac{1}{\sigma^3} = 0$

$-\frac{n}{\sigma} + \frac{\sum(x_i-\mu)^2}{\sigma^3} = 0$

$\hat{\sigma}^2 = \frac{1}{n}\sum(x_i-\hat{\mu})^2$

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

Multivariate Normal (Gaussian) PDFs

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

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

- 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

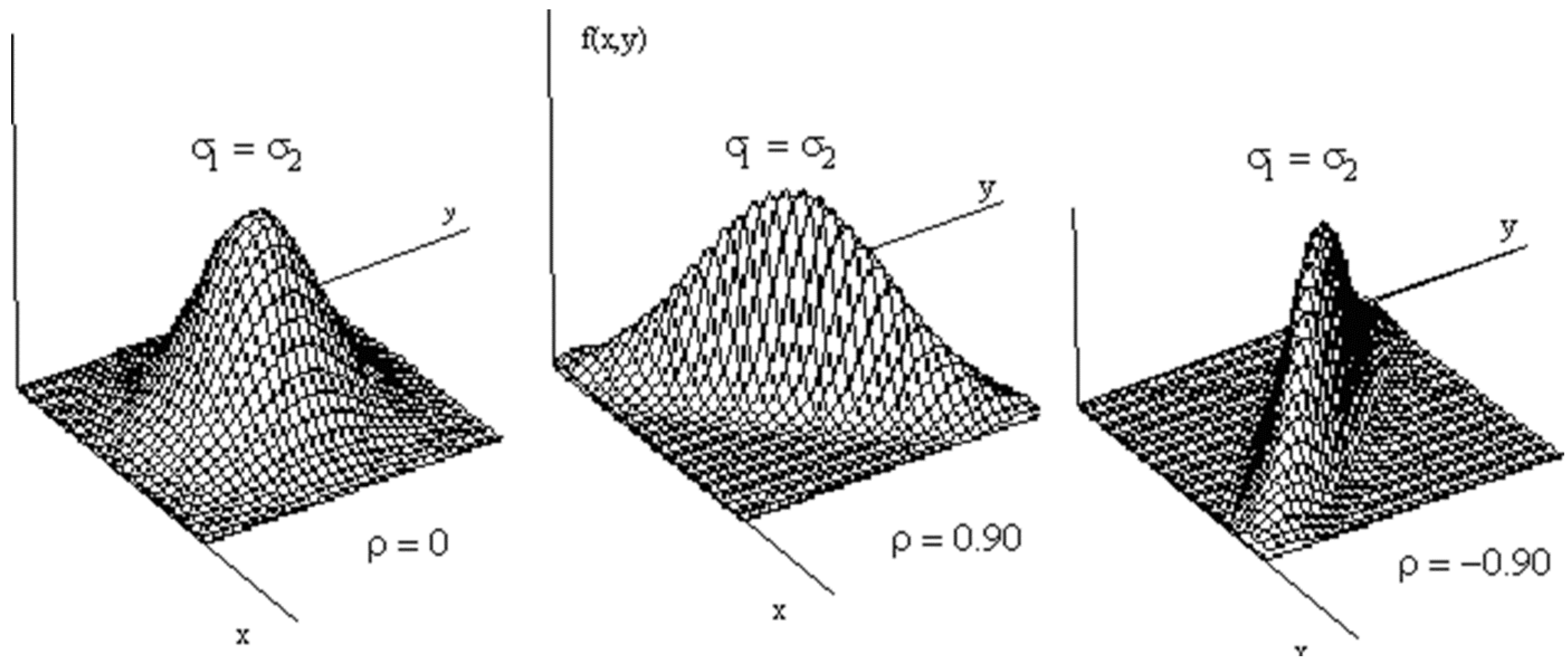
$$f\left(x_1, x_2\right) = \frac{1}{\left(2\pi\right)\left|\Sigma\right|^{1/2}} \mathrm{e}^{-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)}$$

$$\text{with} \quad \vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and}$$

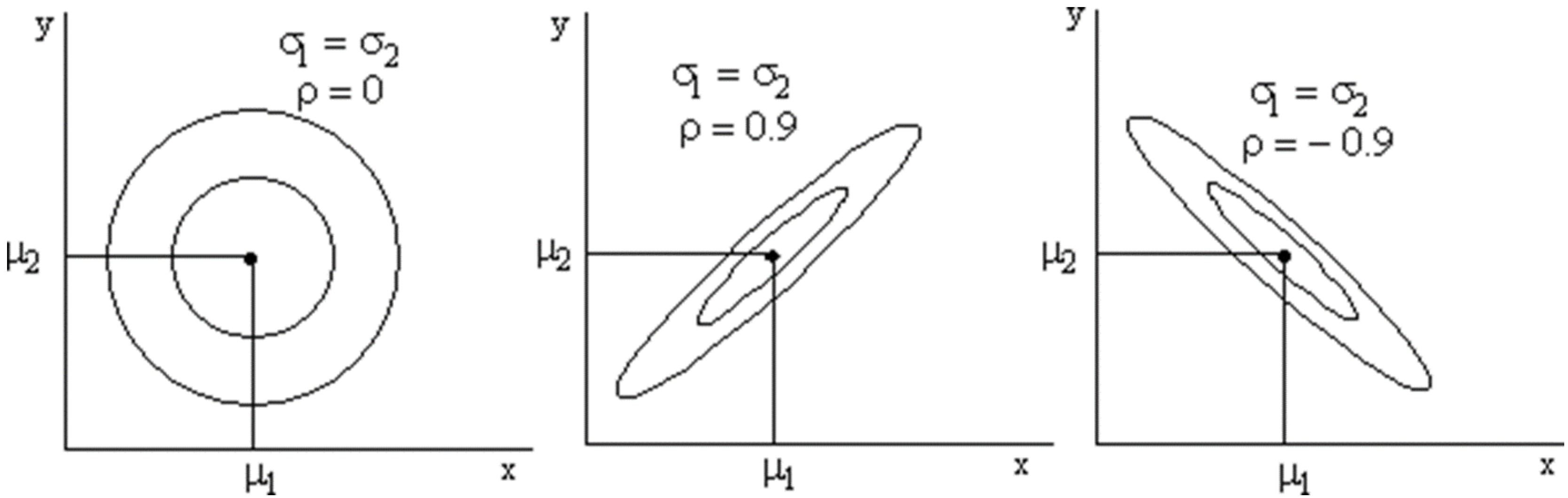
$$\underset{2\times 2}{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}_{2\times 2}$$

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

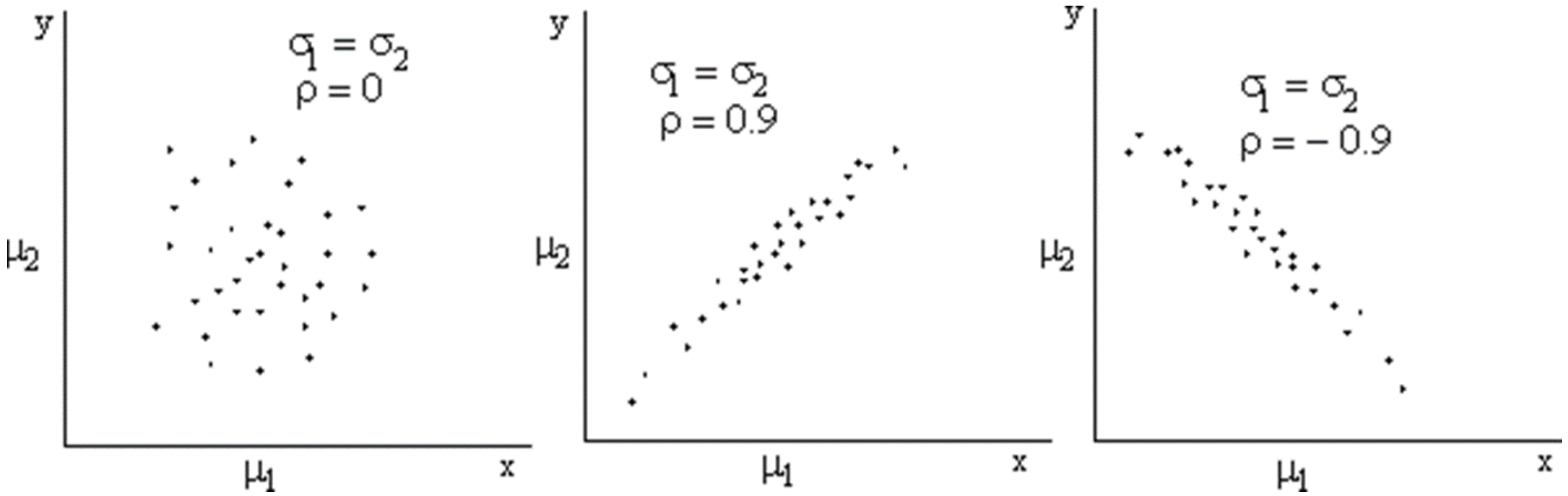
Surface Plots of the bivariate Normal distribution



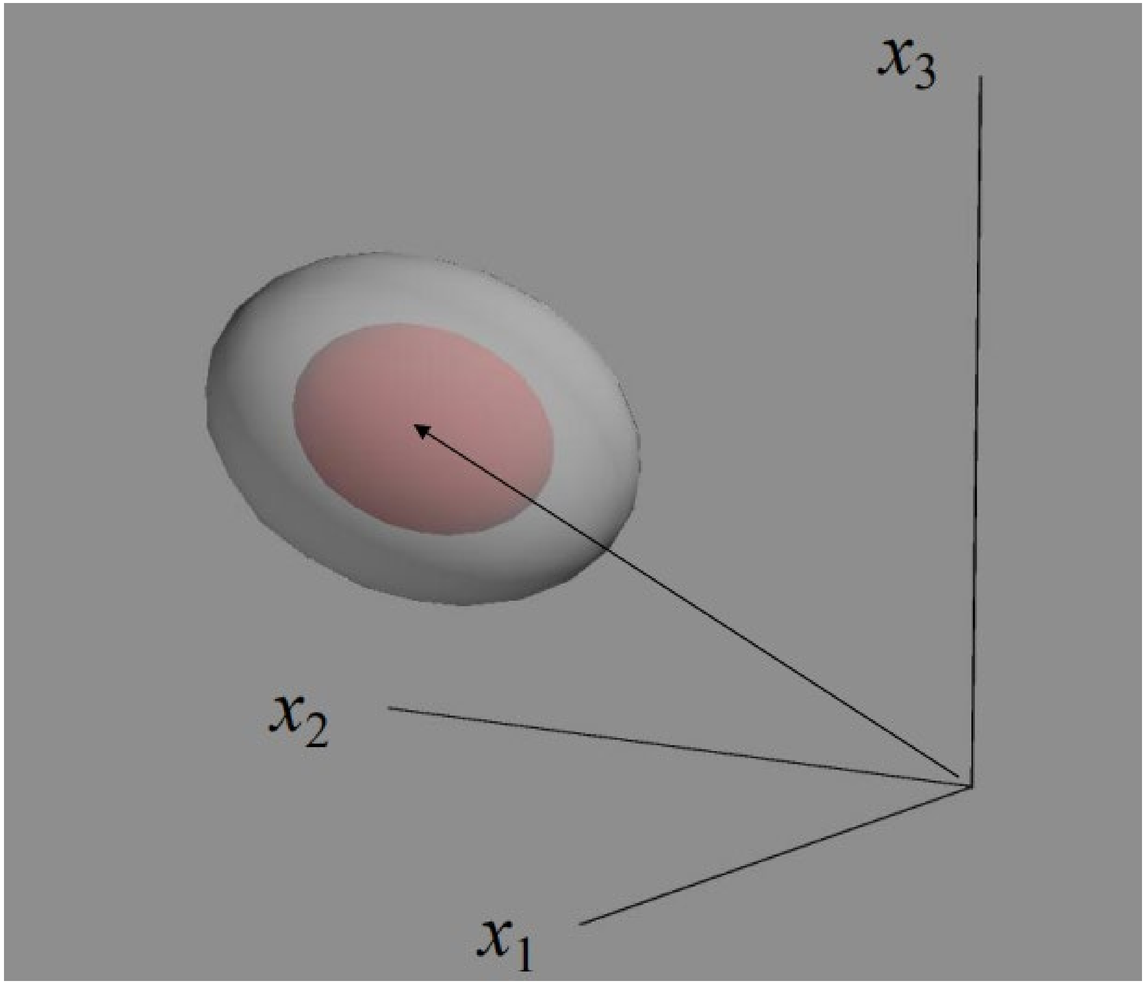
Contour Plots of the bivariate Normal distribution



Scatter Plots of the bivariate Normal distribution

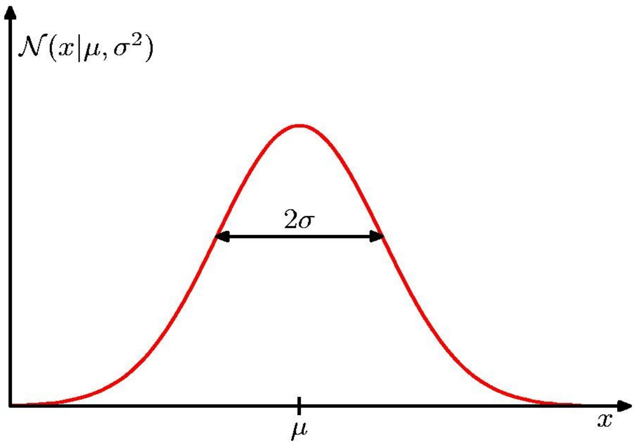


Trivariate Normal distribution



Use MLE to estimate 1-D Gaussian

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



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

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

Use MLE to estimate p-D Gaussian

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

$$\vec{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \qquad \Sigma_{p \times p} = \begin{bmatrix} var(X_1) & \ldots & cov(X_1, X_p) \\ \vdots & \ddots & \vdots \\ cov(X_p, X_1) & \ldots & var(X_p) \end{bmatrix}$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

Today

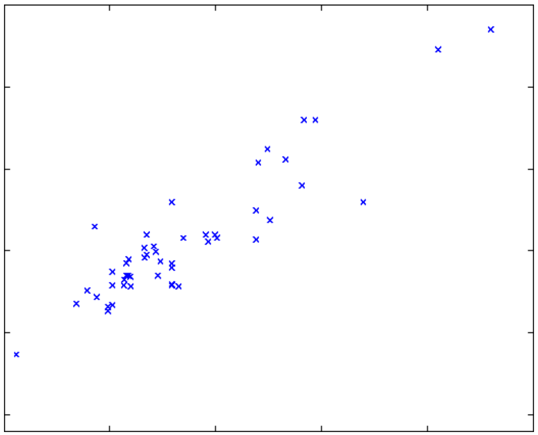
- Basic MLE
- MLE for Discrete RV
- MLE for Continuous RV (Gaussian)
- MLE connects to Normal Equation of LR
- Extra: Properties about Mean and Variance

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:

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

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

- Now assume that ε follows a Gaussian $N(0, \sigma^2)$, then we have:

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

DETOUR: Probabilistic Interpretation of Linear Regression

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

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

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

- We can learn θ by maximizing the likelihood of generating the observed samples

MLE connects to Normal Equation of LR

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

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

$$\Big\Downarrow \ \mathrm{argmax}\, l(\theta) \Rightarrow \mathrm{argmin}\, J(\theta)$$

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

References

- https://qiyanjun.github.io/2019f-UVA-CS6316-MachineLearning/
- Prof. Andrew Moore’s review tutorial
- Prof. Nando de Freitas’s review slides
- Prof. Carlos Guestrin recitation slides

Today

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

Properties

- Correlation:

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

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

Properties

- Mean

$$E(X+Y) = E(X) + E(Y)$$
$$E(aX) = aE(X)$$

  - If X and Y are independent,

$$E(XY) = E(X)E(Y)$$

- Variance

$$V(aX+b) = a^2V(X)$$

  - If X and Y are independent,

$$V(X+Y) = V(X) + V(Y)$$

Properties

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

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

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

$$E(Y) = E[E(Y|X)] = \int E(Y|X = x) p_x(x) dx$$

- The law of Total Variance:

$$Var(Y) = Var[E(Y|X)] + E[Var(Y|X)]$$

Thanks for listening