

PROBABILITY REVIEW(?)

Predicting amount of rainfall



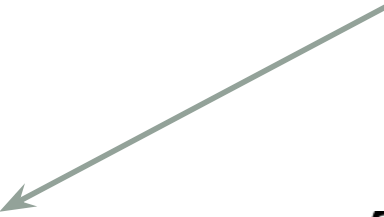
(Linear) Regression

- $h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5$

- θ s are the parameter (or weights)

Assume x_0 is always 1

- We can rewrite

$$h_{\theta}(x) = \sum_{i=0}^n \theta_i x_i = \theta^T \mathbf{x}$$


- Notation: vectors are bolded
- Notation: vectors are column vectors



LMS regression with gradient descent

$$\frac{\partial J}{\partial \theta_j} = -\sum_{i=1}^m (y_i - \theta^T \mathbf{x}_i) x_i^{(j)}$$

$$\theta_j \Leftarrow \theta_j + r \sum_{i=1}^m (y_i - \theta^T \mathbf{x}_i) x_i^{(j)}$$

Interpretation?

Logistic Regression

- Pass $\theta^T \mathbf{x}$ through the logistic function

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Logistic Regression update rule

$$\theta_j \Leftarrow \theta_j - r \sum_{i=1}^m (y_i - h_{\theta}(x_i)) x_i^{(j)}$$

Update rule for linear regression

$$\theta_j \Leftarrow \theta_j - r \sum_{i=1}^m (y_i - \theta^T \mathbf{x}_i) x_i^{(j)}$$

What is Probability?

- Frequentist

Probability = rate of occurrence in a large number of trials

- Bayesian

Probability = uncertainty in your knowledge of the world



Bayesian vs Frequentist



Bayesian vs Frequentist

- Toss a coin
- Frequentist
 - $P(\text{head}) = \theta$, $\theta = \text{\#heads}/\text{\#tosses}$
- Bayesian
 - $P(\text{head}) = \theta$, $\theta \sim U(0.6, 1.0)$
 - Parameters of distributions can now have probabilities
 - Bayesian interpretation can give prior knowledge to the phenomena – subjective view of the world
 - Prior knowledge can be updated according to the observed frequency

Bayesian statistics

- Coin with $P(\text{head}) = p$
- Observed frequency of heads $\hat{p} = \text{\#heads}/\text{\#n}$
- In Bayesian view, we can talk about $P(p \mid \hat{p})$ by using Bayes's rule

$$P(p|\hat{p}) = \frac{P(\hat{p}|p)P(p)}{P(\hat{p})}$$

Prior probability



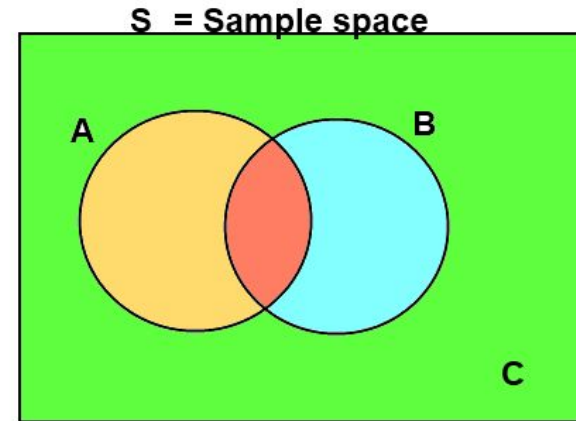
Important concepts

- Conditional probability
 - Independence
- Bayes' Rule
- Expected Value and Variance
- CDFs
- Sum of RVs
- Gaussian Random Variable
 - Multivariate Gaussian

Conditional probability

- $P(A|B)$ probability of A given B has occurred

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



- A student posts a facebook status after finishing Pattern Recognition homework
- $P(\text{he is happy})$
- $P(\text{he is happy} \mid \text{the post starts with “\#\$@\#\$!@\#\$”})$

Independence

- Two events are independent (statistically independent or stochastically independent) if the occurrence of one does not affect the probability of occurrence of the other.

$$P(A \cap B) = P(A)P(B) \Leftrightarrow P(B) = P(B \mid A)$$

- $P(\text{he is happy} \mid \text{His friend posted a cat picture on instragram})$

Bayes' Rule (Bayes's theorem or Bayes' law)

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

Usefulness: We can find $P(A|B)$ from $P(B|A)$ and vice versa

Expected value

- Expected value

$$E[x] = \int_{-\infty}^{\infty} xp(x)dx$$

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)p(x)dx$$

- Variance (σ^2) (Standard Deviation = σ)

$$Var[x] = E[(x - E[x])^2] = \sigma^2 = \int_{-\infty}^{\infty} (x - E[x])^2 p(x)dx$$

$$E[(x - E[x])^2] = E[x^2] - (E[x])^2$$

Expected Value notes

- It's a weighted sum.
- Something can have a high expected value but low probability of occurring
- Lottery: $P(\text{win}) = 10^{-20}$, winner gets 10^{30}
- $P(\text{loss}) = 1 - P(\text{win})$, loser gets -10
- $E(\text{Lottery earnings}) = 10^{-20}10^{30} + (1 - P(\text{win}))(-10)$
- $= 10^{10} - 10$
- Humans are not good at gauging probability at extreme points

Expected value and Variance properties

- $E[a] = a$; a is a constant.
- $E[aX+b] = aE[X]+b$
- $E[X+Y] = E[X]+E[Y]$
- $\text{Var}[a] = 0$
- $\text{Var}[aX+b] = a^2\text{Var}[X]$

Conditional Expected Value

$$E[x | A] = \int_{-\infty}^{\infty} xp(x | A)dx$$

$$E[g(x) | A] = \int_{-\infty}^{\infty} g(x)p(x | A)dx$$

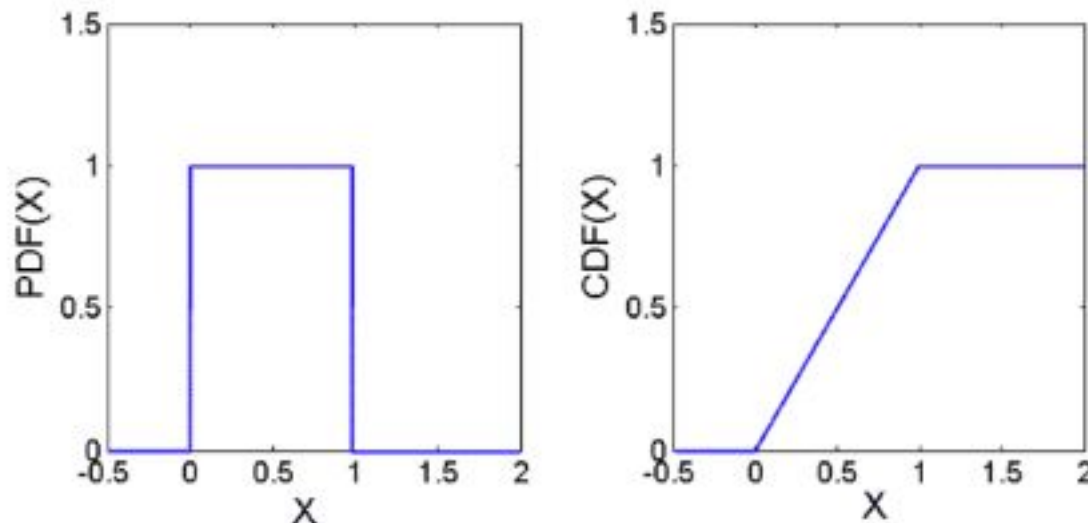
Cumulative Distribution Functions

CDFs

- Probability that the RV is less than a certain amount

$$F_X(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} p(x)dx$$

- CDF is the integral of PDF. Differentiating CDF wrt x gives the PDF



Joint distributions

- If we want to monitor how two events are jointly occurring, we consider the joint distribution $p_{X,Y}(x,y)$
- $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ if x and y are independent

$$P(A) = \int \int_A p_{XY}(x, y) dx dy$$

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, y) dy$$

$$p_Y(y) = \int_{-\infty}^{\infty} p_{XY}(x, y) dx$$

Sum of Random variables

- $Z = Y + X$
- What is the pdf of Z ? Where Y and X continuous RVs

$$p_{X+Y}(z) = (p_X * p_Y)(z) = (p_Y * p_X)(z)$$

Central Limit Theorem (CLT)

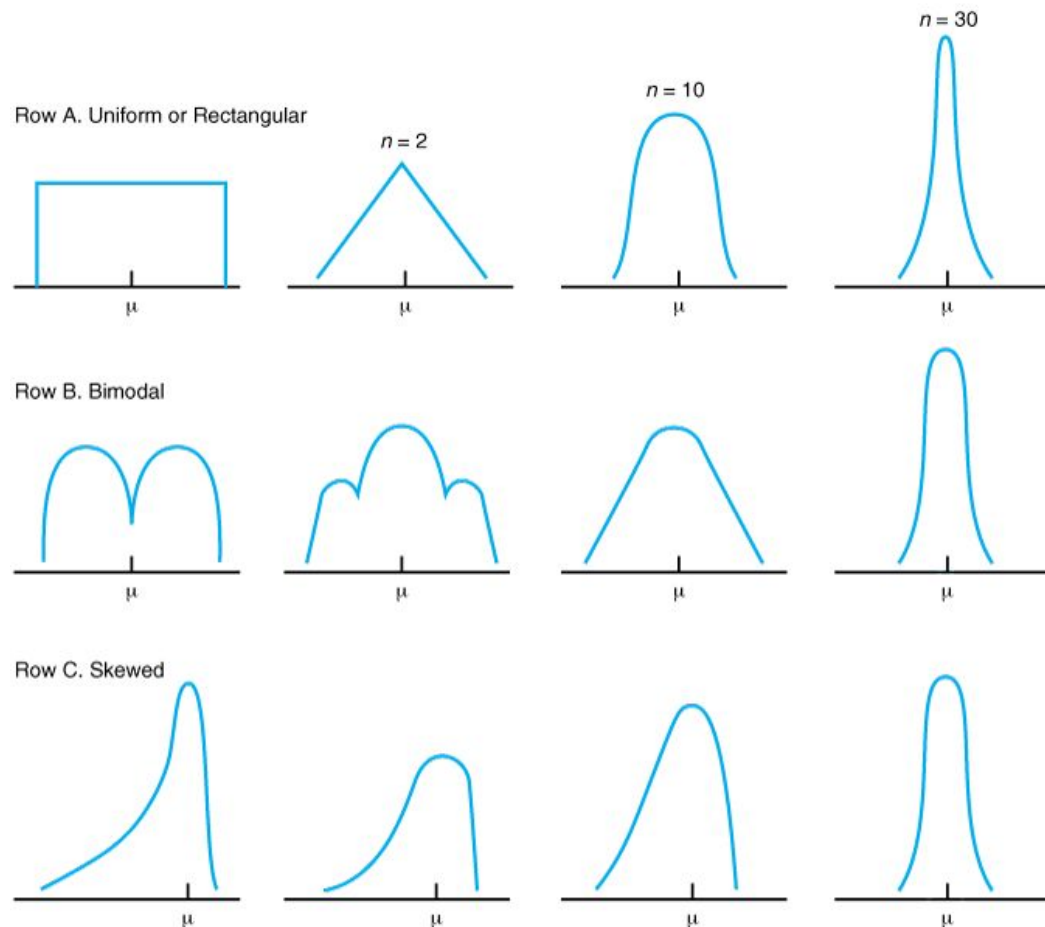
- Suppose X_1, X_2, \dots is a sequence of iid (independent and identically distributed) RVs. As n approaches infinity the sum of the sequence converge in distribution to a Normal distribution

$$\sqrt{n} \left(\left(\frac{1}{n} \sum_{i=1}^n X_i \right) - \mu \right) \xrightarrow{d} N(0, \sigma^2)$$

- Other variants of CLT exists, without the dependence or identically distributed assumption

CLT implications

- A sum of RVs tends to become Normally distributed very quickly



Gaussian distribution (normal distribution)

- X is normal (Gaussian): $X \sim N(\mu, \sigma^2)$

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

$$E[x] = \mu$$

$$\text{Var}[x] = \sigma^2$$

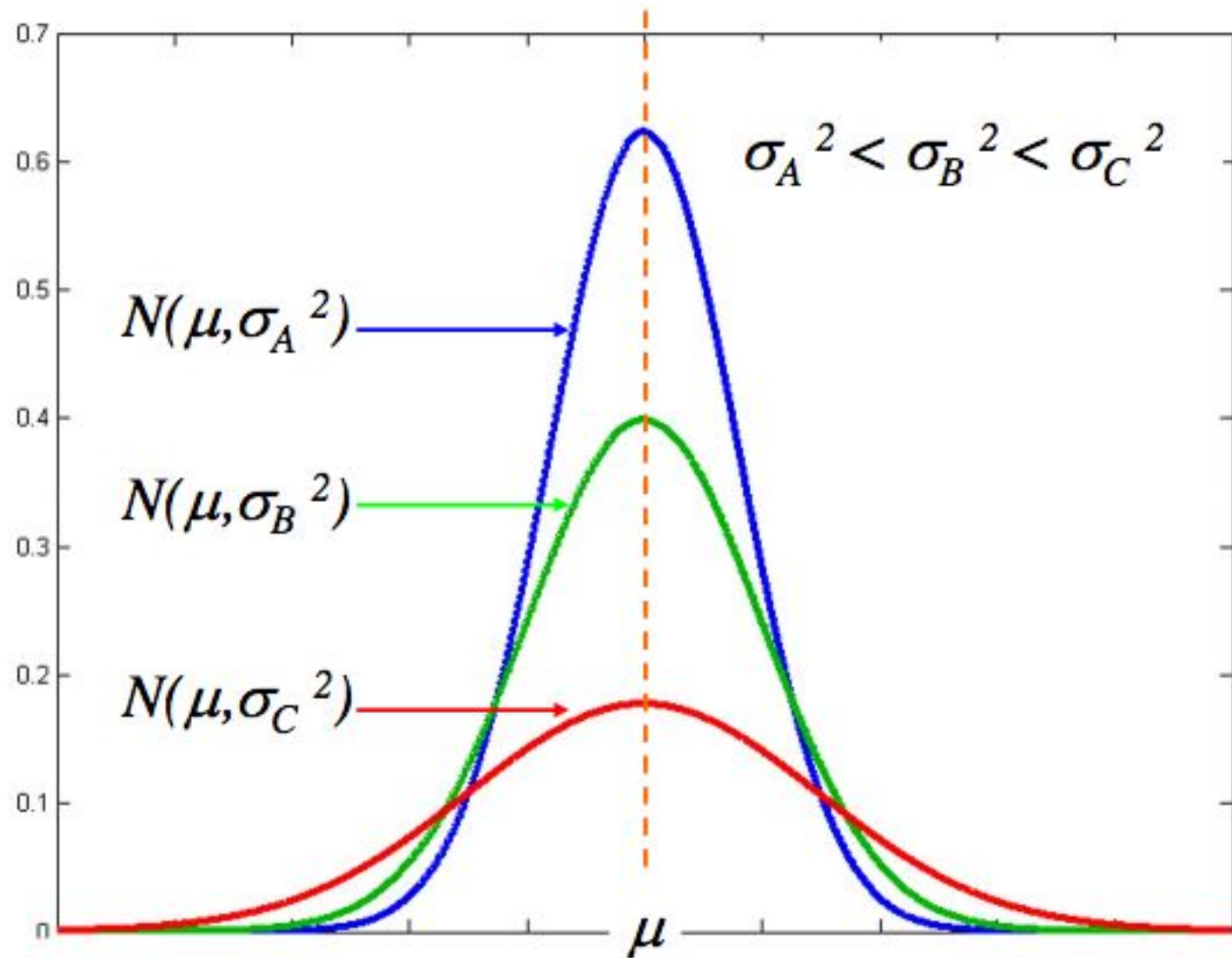
- X is Standard normal (Standard Gaussian):
 $X \sim N(0,1)$ when $\mu=0, \sigma^2=1$

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

$$E[x] = 0$$

$$\text{Var}[x] = 1$$

Gaussian pdf



Linear transformation of Gaussian RV

- Normality is preserved by linear transformation. Calculation involving the normal variable is usually done in terms of standard normal.
- Let $Y=aX+b$,
if $X \sim N(\mu, \sigma^2) \rightarrow Y \sim N(a\mu+b, a^2 \sigma^2)$
- Let $Z=(X-\mu)/\sigma$,
if $X \sim N(\mu, \sigma^2) \rightarrow Z \sim N(0,1) : \text{Standard Normal}$

Can you prove this?

Summation of 2 Gaussian RVs

- X mean m_1 variance σ_1^2
 - Y mean m_2 variance σ_2^2
 - X and Y are independent
-
- X+Y is normally distributed with mean m_1+m_2 variance $\sigma_1^2+\sigma_2^2$

Expectation of multivariate distributions

$$E[g(X_1, X_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_1, x_2) p_{X_1, X_2}(x_1, x_2) dx_1 dx_2$$

$$E[g(X_1)h(X_2)] = E[g(X_1)]E[h(X_2)]$$

If X_1 and X_2 independent

Covariance of multivariate distributions

- $\text{cov}(X_1, X_2) = E[(X_1 - m_1)(X_2 - m_2)]$
- $\text{cov}(X_1, X_2) = E[(X_1)(X_2)] - m_1 m_2$
- Covariance with itself is just the Variance
- Correlation

$$\rho = \frac{\text{cov}(X_1, X_2)}{\sqrt{V(X_1)V(X_2)}}$$

Covariance matrix

- Given a set of RVs, $X_1 X_2 \dots X_n$
- The covariance matrix is a matrix which has the covariance of the i and j RV in position (i,j)

$$\Sigma = \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & E[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & E[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & E[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

Understanding the Covariance matrix

$$= \begin{matrix} \begin{matrix} \text{A} \\ \text{C} \end{matrix} & \begin{bmatrix} \text{Cov}(X, X) & \text{Cov}(Y, X) \\ \text{Cov}(X, Y) & \text{Cov}(Y, Y) \end{bmatrix} & \begin{matrix} \text{B} \\ \text{D} \end{matrix} \end{matrix}$$

Which statements are true?

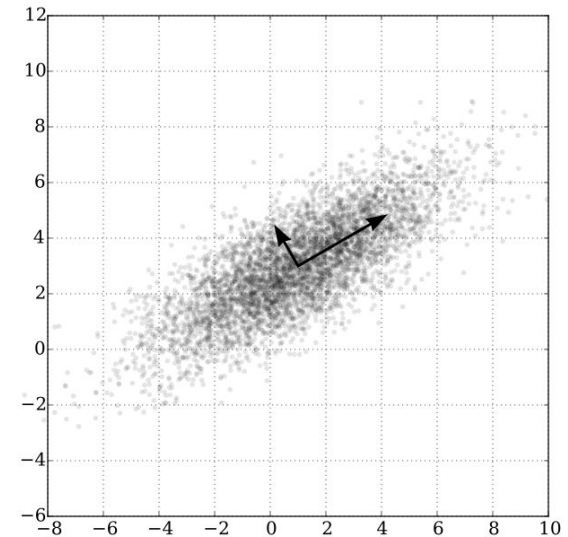
$$C = B$$

$$B < 0$$

$$D < 0$$

$$A < D$$

$$A > B$$



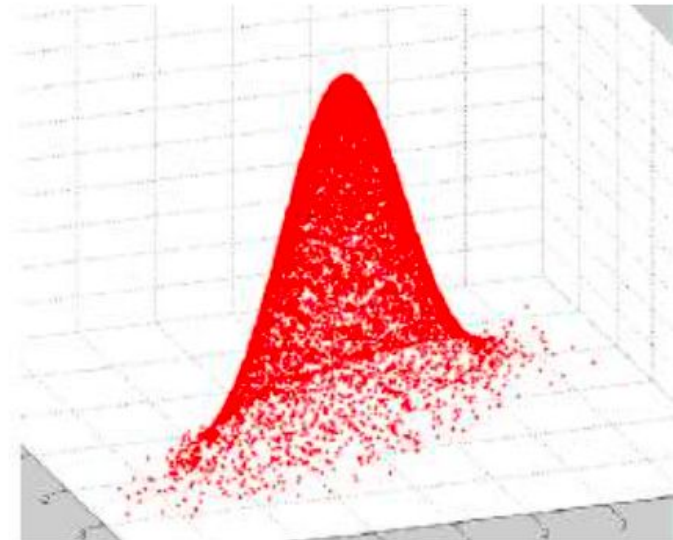
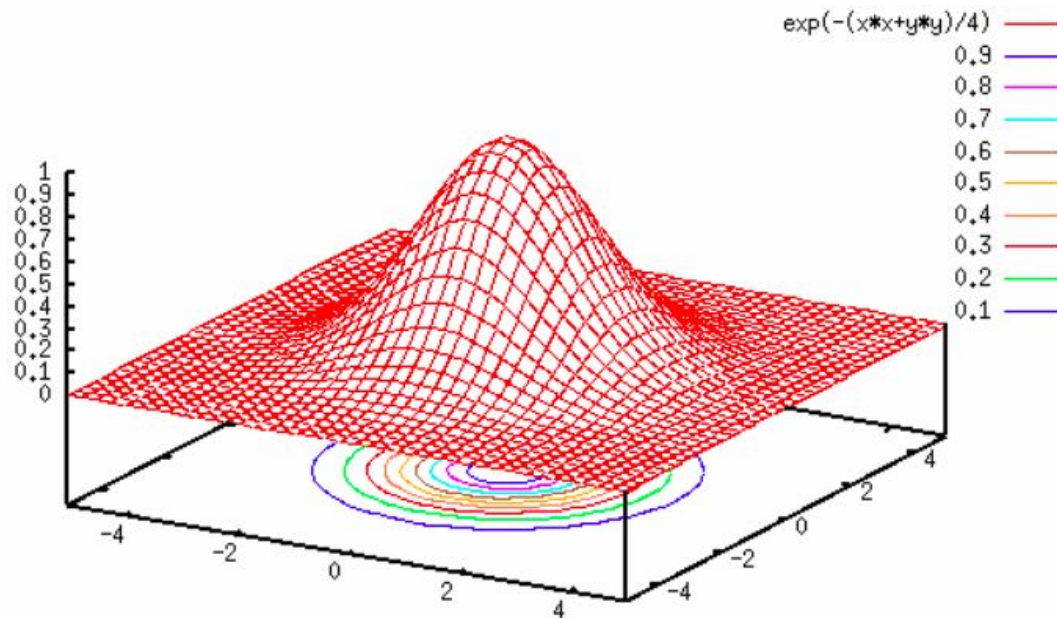
Covariance matrix observations

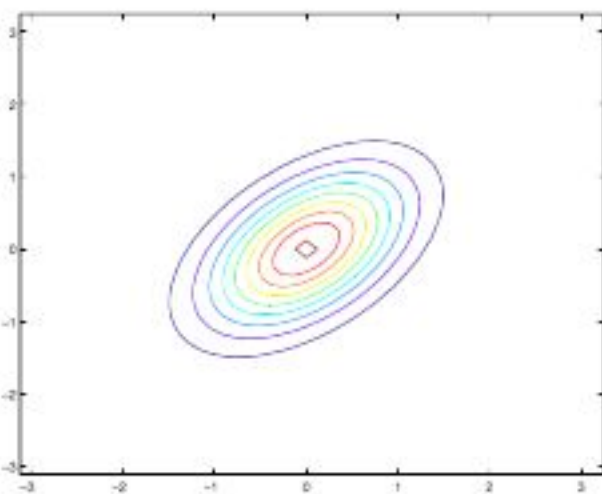
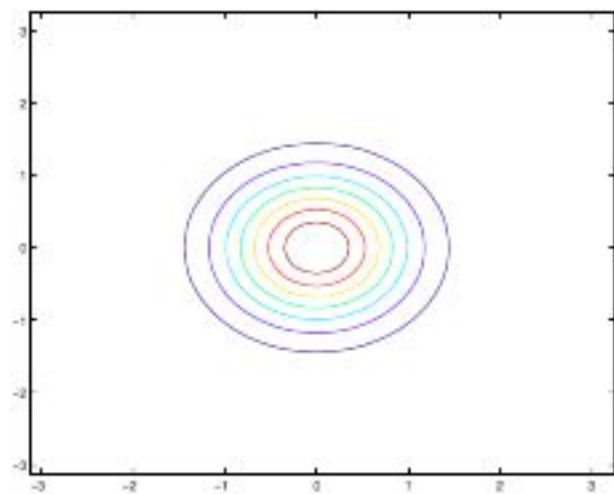
- $\Sigma = \Sigma^T$
- If the covariance matrix is diagonal, all RVs are mutually independent.
- Covariance matrix is positive-semidefinite
 - Every positive definite matrix is invertible

Multivariate Gaussian distribution

- Put $X_1, X_2, X_3 \dots X_n$ into a vector \mathbf{x}

$$p(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^t \Sigma^{-1}(\mathbf{x} - \mu)\right]$$



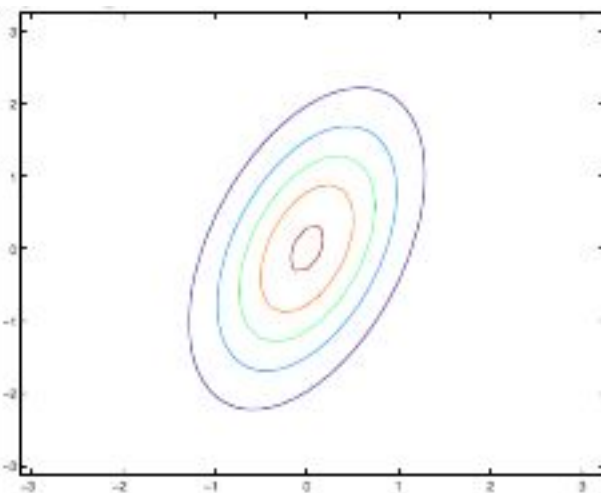
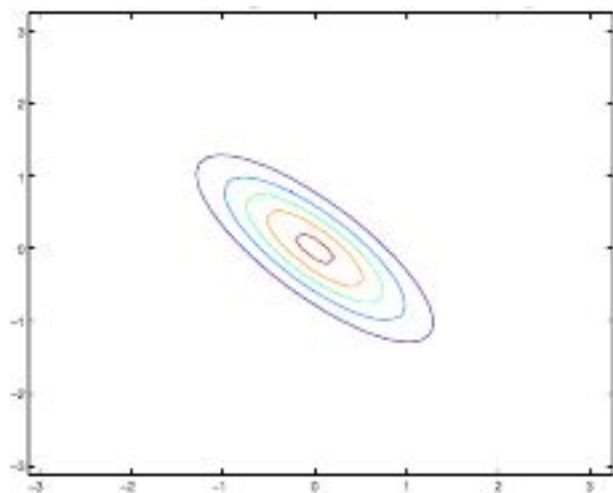


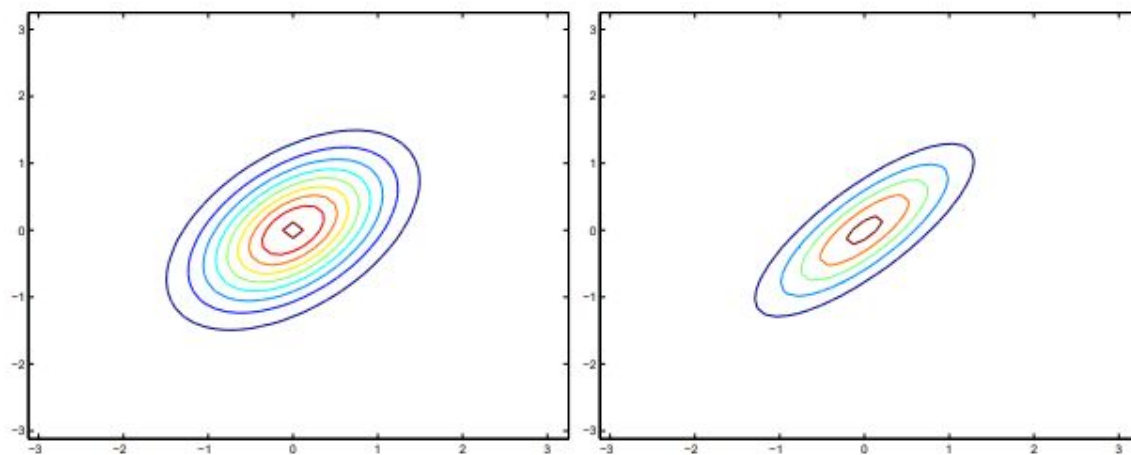
$$^1 \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$

$$^2 \Sigma = \begin{bmatrix} 3 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

$$^3 \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$^4 \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$





1

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

3

$$\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

Affine transformation of multivariate Gaussians

- $\mathbf{y} = \mathbf{Ax} + \mathbf{b}$, Assuming A has full rank (invertible)

$$\mathbf{x} \sim N(\mu, \Sigma)$$

$$\mathbf{y} \sim N(A\mu + b, A\Sigma A^T)$$

Important concepts

- Conditional probability
 - Independence
- Bayes' Rule
- Expected Value and Variance
- CDFs
- Sum of RVs
 - CLT
- Gaussian Random Variable
 - Multivariate Gaussian

Distribution parameter estimation

- [illegible]

Linear Regression Revisit

- $h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5$
- θ s are the parameter (or weights)
- We can rewrite

$$h_{\theta}(x) = \sum_{i=0}^n \theta_i x_i = \theta^T \mathbf{x}$$

- Notation: vectors are bolded
- Notation: vectors are column vectors



Probabilistic Interpretation of linear regression

- Real world data is our model plus some error term
 - Noise in the data
 - Something that we do not model (features we are missing)
- Let's assume the error is normally distributed with mean zero and variance σ^2
 - Why Gaussian?
 - Why saying mean is zero is a valid assumption?

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

Probabilistic view of Linear regression

- Find θ
- Maximize Likelihood of seeing x and y in training
- From our assumption we know that

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

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

Error term is normally distributed with mean 0 and variance σ^2

Maximizing Likelihood

What is the assumption here?
Is it accurate?

- Max $L(\theta) = \prod_{i=1}^m p(y_i | \mathbf{x}_i; \theta)$
- We use the log likelihood instead $\log(L(\theta)) = l(\theta)$

From our previous lecture

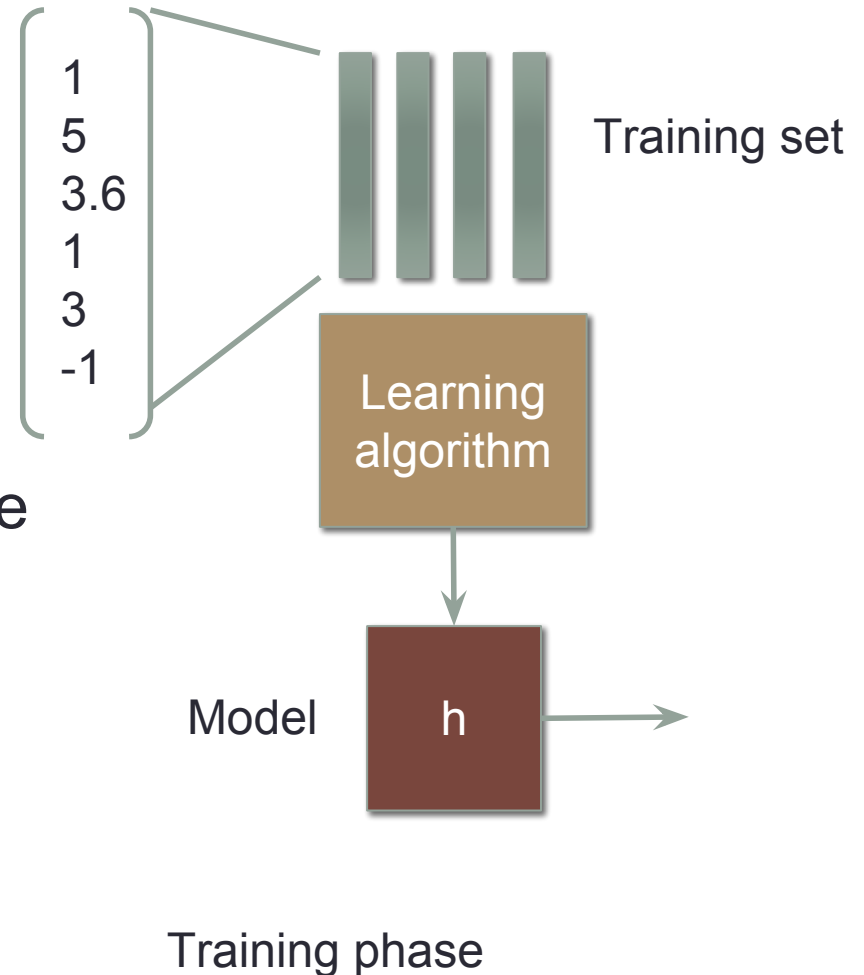
$$\text{Min } J(\theta) = \frac{1}{m} \sum_{i=1}^m (y_i - \theta^T \mathbf{x}_i)^2$$

Mean square error solution and MLE solution

- Turns out MLE and MSE gets to the same solution
 - This justifies our choice of MSE as the Loss for linear regression
 - This does not mean MSE is the best Loss for regression, but you can at least justify it with a probabilistic reasoning
- Note how our choice of variance σ^2 falls out of the maximization, so this derivation is true regardless of which assumption for variance is.

Flood or no flood

- What would be the output?
- $y = 0$ if not flooded
- $y = 1$ if flooded
- Anything in between is a score for how likely it is to flood

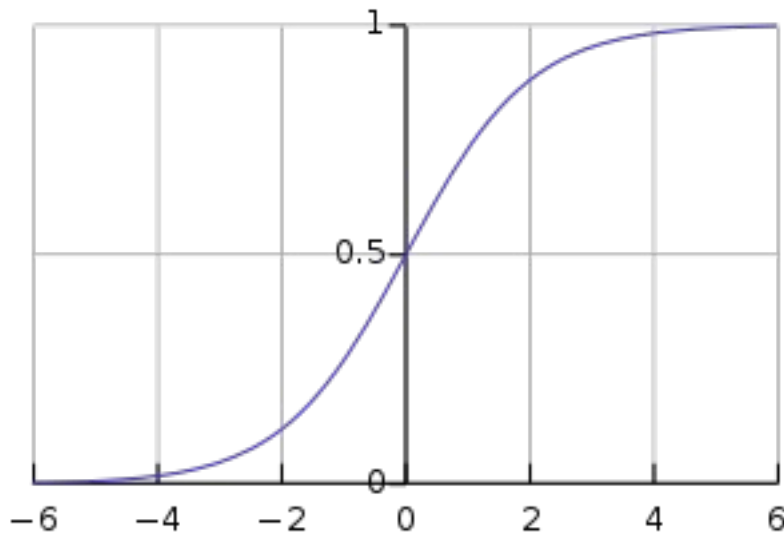


Can we use regression?

- Yes
- $h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5$
- But
- What does it mean when h is higher than 1?
- Can h be negative? What does it mean to have a negative flood value?

Logistic function

- Let's force h to be between 0 and 1 somehow
- Introducing the logistic function (sigmoid function)



$$\begin{aligned} f(x) &= \frac{1}{1 + e^{-x}} \\ &= \frac{e^x}{1 + e^x} \end{aligned}$$

Logistic Regression

- Pass $\theta^T \mathbf{x}$ through the logistic function

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Loss function?

- MSE error no longer a good candidate
- Let's turn to use probabilistic argument for logistic regression

Logistic Function derivative

The derivative has a nice property by design.

This is also why many algorithm we'll learn later in class also uses the logistic function

$$\begin{aligned} g'(z) &= \frac{d}{dz} \frac{1}{1 + e^{-z}} \\ &= \frac{1}{(1 + e^{-z})^2} (e^{-z}) \\ &= \frac{1}{(1 + e^{-z})} \cdot \left(1 - \frac{1}{(1 + e^{-z})} \right) \\ &= g(z)(1 - g(z)). \end{aligned}$$

Probabilistic view of Logistic Regression

- Let's assume, we'll classify as 1 with probability in accordance to the output of

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$P(y = 1 \mid x; \theta) = h_{\theta}(x)$$

$$P(y = 0 \mid x; \theta) = 1 - h_{\theta}(x)$$

or

$$p(y \mid x; \theta) = (h_{\theta}(x))^y (1 - h_{\theta}(x))^{1-y}$$

Maximizing log likelihood

$$p(y \mid x; \theta) = (h_{\theta}(x))^y (1 - h_{\theta}(x))^{1-y}$$

$$\begin{aligned} g'(z) &= \frac{d}{dz} \frac{1}{1 + e^{-z}} \\ &= \frac{1}{(1 + e^{-z})^2} (e^{-z}) \\ &= \frac{1}{(1 + e^{-z})} \cdot \left(1 - \frac{1}{(1 + e^{-z})}\right) \\ &= g(z)(1 - g(z)). \end{aligned}$$

$$\theta_j \Leftarrow \theta_j + r \sum_{i=1}^m (y_i - h_{\theta}(x_i)) x_i^{(j)}$$

Logistic Regression update rule

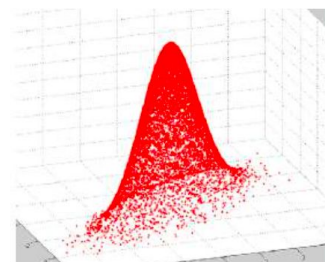
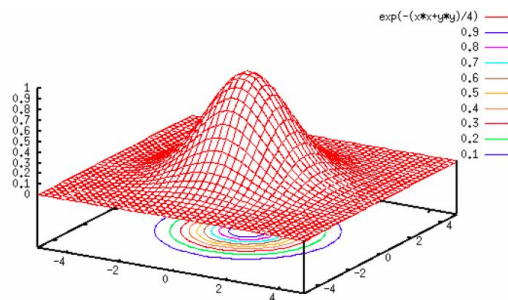
$$\theta_j \Leftarrow \theta_j + r \sum_{i=1}^m (y_i - h_{\theta}(x_i)) x_i^{(j)}$$

Update rule for linear regression

$$\theta_j \Leftarrow \theta_j + r \sum_{i=1}^m (y_i - \theta^T \mathbf{x}_i) x_i^{(j)}$$

Summary

- Conditional probability
 - Independence
- Bayes' Rule
- Expected Value and Variance
- CDFs
- Sum of RVs
 - CLT
- Gaussian Random Variable
 - Multivariate Gaussian
- Probabilistic view for Regression setups



Next time

- HW1 due Monday, Quiz 1 starts 13:10 ends 13:20. HW2 out
- MLE on more things!
- Maximum A Posteriori Estimation (MAP estimate)
- Classification based on probabilistic models

Homework

Can I just use scikit-learn?

No

You can use pandas and numpy for this homework



François Chollet  @fchollet · Aug 25

A popular quote goes "if you can't explain it in simple terms, you don't understand it well enough" (often incorrectly attributed to Einstein or Feynman).

I think a more accurate take is: "if you can't explain it in arbitrarily precise terms, you don't understand it well enough"



19



101



458



François Chollet 

@fchollet

Follow

In particular, if you understand something clearly, you should be able to describe it in precise algorithmic terms to a computer: you should be able to implement it from scratch (as a simulation, as a framework, etc).

Creator of Keras

One button machines

- Machine learning as a tool for non-experts
- Can a non-expert just provide the data and let the machine decide how to proceed

DataRobot

PRODUCT ▾

SOLUTIONS ▾

EDUCATION ▾

ABOUT ▾

WE'RE HIRING! ▾



CONTACT US

- 1 Upload your data
- 2 Select the target variable
- 3 Build 100s of models in one click
- 4 Explore top models and get insights
- 5 Deploy best model and make predictions

Summary

What would you like to predict?

rea

readmitted

Feature name	Var type	Unique	Missing	...
race	Categorical	5	221	
gender	Categorical	2	0	
age	Categorical	10	0	

Reinforcement Learning for Model Selection

- Tuning a network takes time
- Let machine learning learn how to tune a network
- Matches or outperforms ML experts performance

