PROBABILITY REVIEW(?)

Predicting amount of rainfall

https://esan108.com/%E0%B8%9E%E0%B8%A3%E0%B8%B0%E0%B9%82%E0%B8%84%E0%B8%81%E0%B8%B4%E0%B8%99%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3-%E0%B8%AB%E0%B8%A1%E0%B8%B2%E0%B8%A2%E0%B8%96%E0%B8%B6%E0%B8%87.html

(Linear) Regression

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

Assume x_0 is always 1

- We can rewrite

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

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

LMS regression with gradient descent

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

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

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

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

Update rule for linear regression

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

What is Probability?

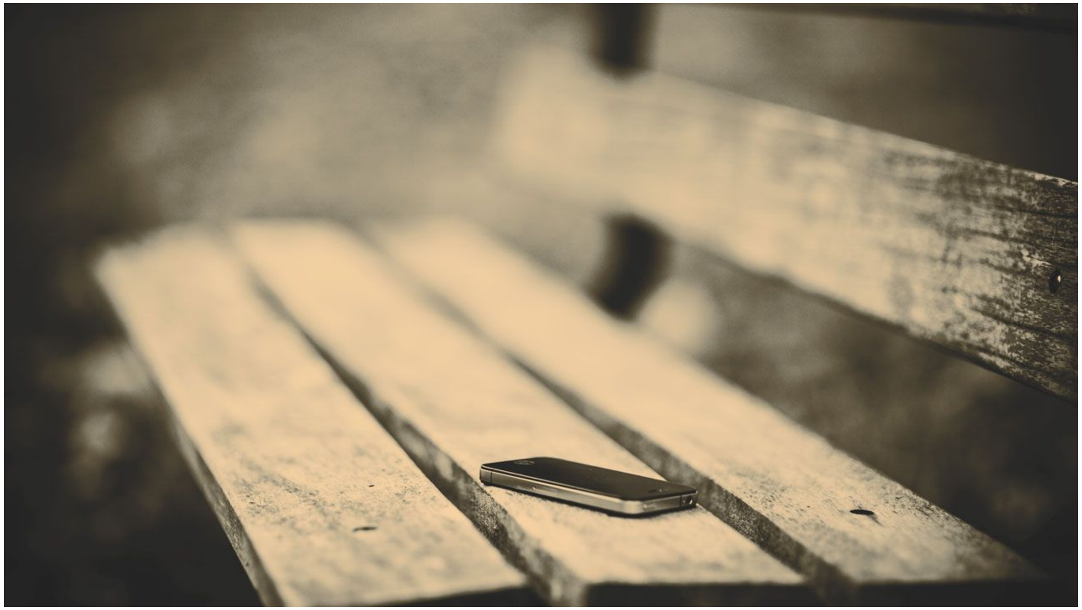
- Frequentist

  Probability = rate of occurrence in a large number of trials

- Bayesian

  Probability = uncertainty in your knowledge of the world

Bayesian vs Frequentist



https://www.gizmodo.com.au/2017/01/how-to-find-a-lost-phone/

Bayesian vs Frequentist

- Toss a coin
- Frequentist
  - $P(head) = \theta$, $\theta = \#heads/\#tosses$
- Bayesian
  - $P(head) = \theta$, $\theta \sim U(0.6, 1.0)$
  - Parameters of distributions can now have probabilities
  - Bayesian interpretation can gives 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(head) = p
- Observed frequency of heads $\hat{p}$ = #heads/#n

- In Bayesian view, we can talk about $P(p \mid \hat{p})$ by using Bayes's rule

Prior probability

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

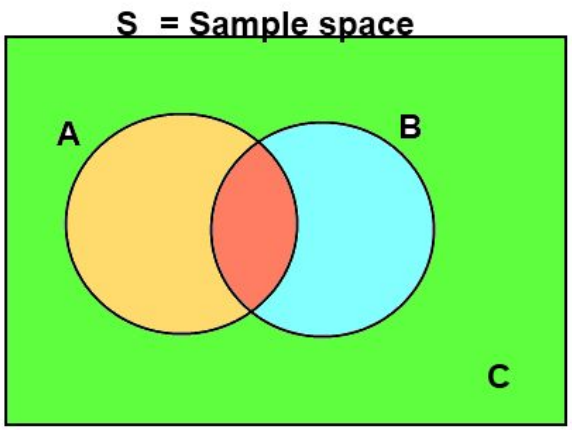
Important concepts

- 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

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



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

https://www.wyzant.com/resources/lessons/math/statistics_and_probability/probability/further_concepts_in_probability

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.

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B) \Leftrightarrow \mathrm{P}(B) = \mathrm{P}(B \mid A)$$

- P(he is happy | His friend posted a cat picture on instragram)

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

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

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

Expected value

- Expected value

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

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

- 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(win) = 10^(-20), winner gets 10^30
- P(loss) = 1-P(win), loser gets -10
- E(Lottery earnings) = 10^(-20)10^30+ (1-P(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.
- $E[aX+b] = aE[X]+b$
- $E[X+Y] = E[X]+E[Y]$
- $Var[a] = 0$
- $Var[aX+b] = a^2Var[X]$

Conditional Expected Value

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

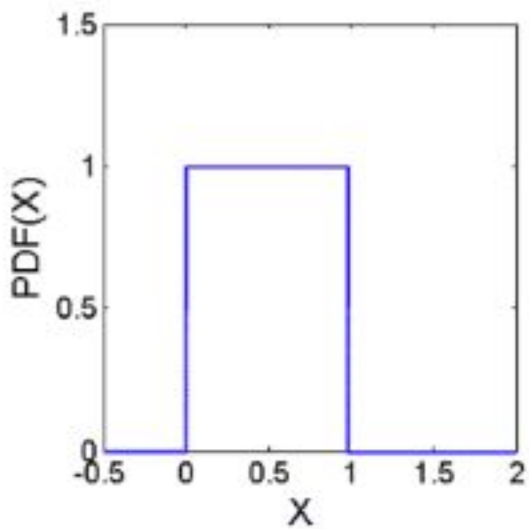
$$E[g(x) \mid A] = \int_{-\infty}^{\infty} g(x)p(x \mid A)dx$$

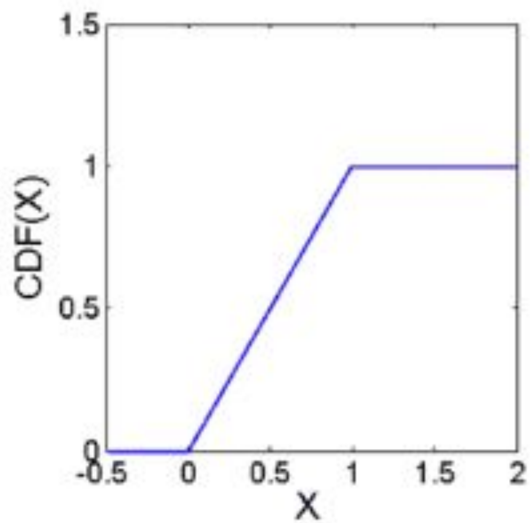
Cumulative Distribution Functions CDFs

- Probability that the RV is less than a certain amount

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

- 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

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

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

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

Sum of Random variables

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

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

Central Limit Theorem (CLT)

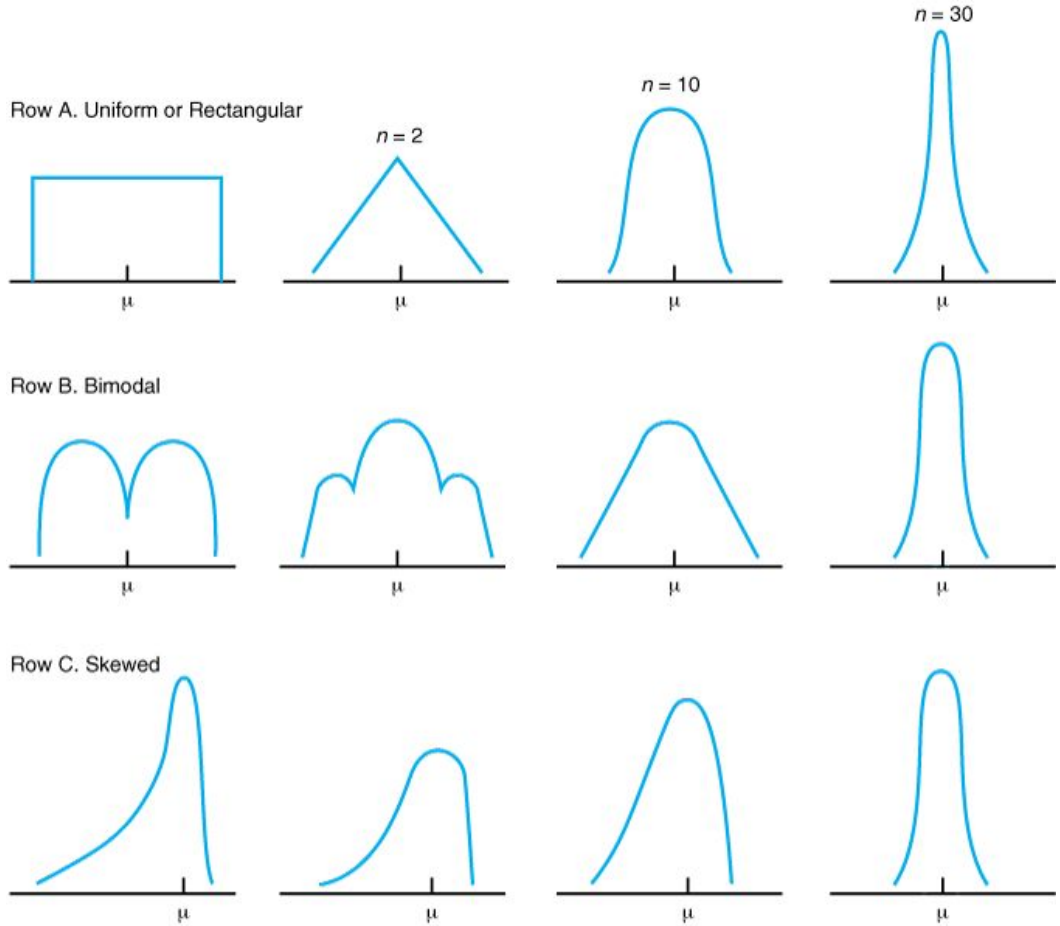
- Suppose X_1,X_2,… is a sequence of iid (independent and identically distributed) RVs. As n approaches infinity the sum of the sequence converge in distribution to a Normal distribution

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

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

CLT implications

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



http://www.people.vcu.edu/~albest/DENS580/DawsonTrapp/Chap4.htm

Gaussian distribution (normal distribution)

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

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

$$E[x] = \mu$$

$$Var[x] = \sigma^2$$

- X is Standard normal (Standard Gaussian):

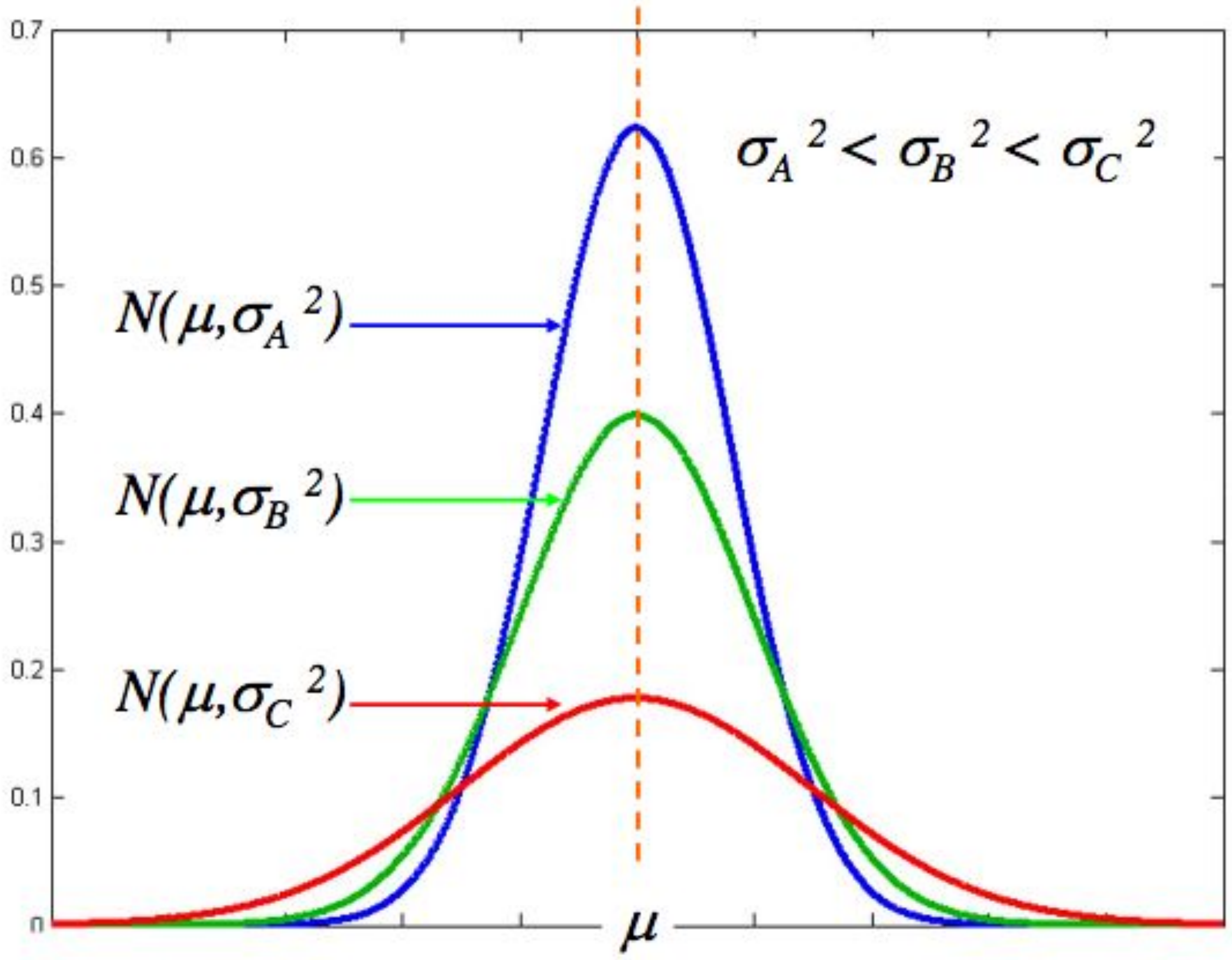
$X \sim N(0,1)$ when $\mu=0$, $\sigma^2=1$

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

$$E[x] = 0$$

$$Var[x] = 1$$

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 $X \sim N(\mu, \sigma^2) \rightarrow Z \sim N(0, 1)$: 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

- $cov(X_1,X_2) = E[(X_1-m_1)(X_2-m_2)]$

- $cov(X_1,X_2) = E[(X_1)(X_2)] - m_1m_2$

- Covariance with itself is just the Variance
- Correlation

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

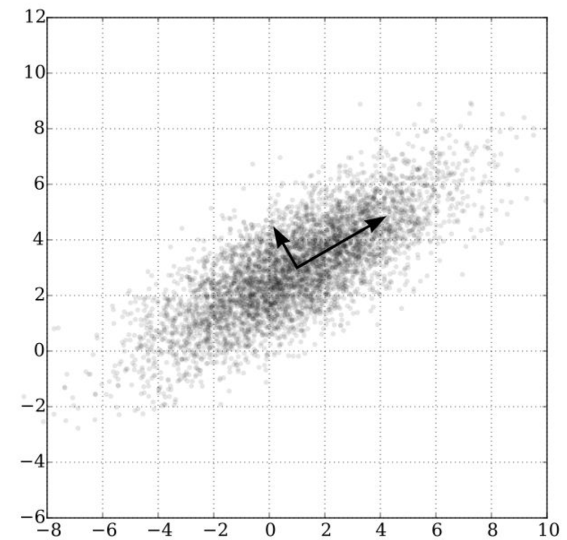
Covariance matrix

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

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

Understanding the Covariance matrix

$$= \begin{array}{lcr} A & & B \\ & \begin{bmatrix} \mathrm{Cov}(X,X) & \mathrm{Cov}(Y,X) \\ \mathrm{Cov}(X,Y) & \mathrm{Cov}(Y,Y) \end{bmatrix} & \\ C & & D \end{array}$$



Which statements are true?

C = B

B < 0

D < 0

A < D

A > B

https://en.wikipedia.org/wiki/Covariance_matrix

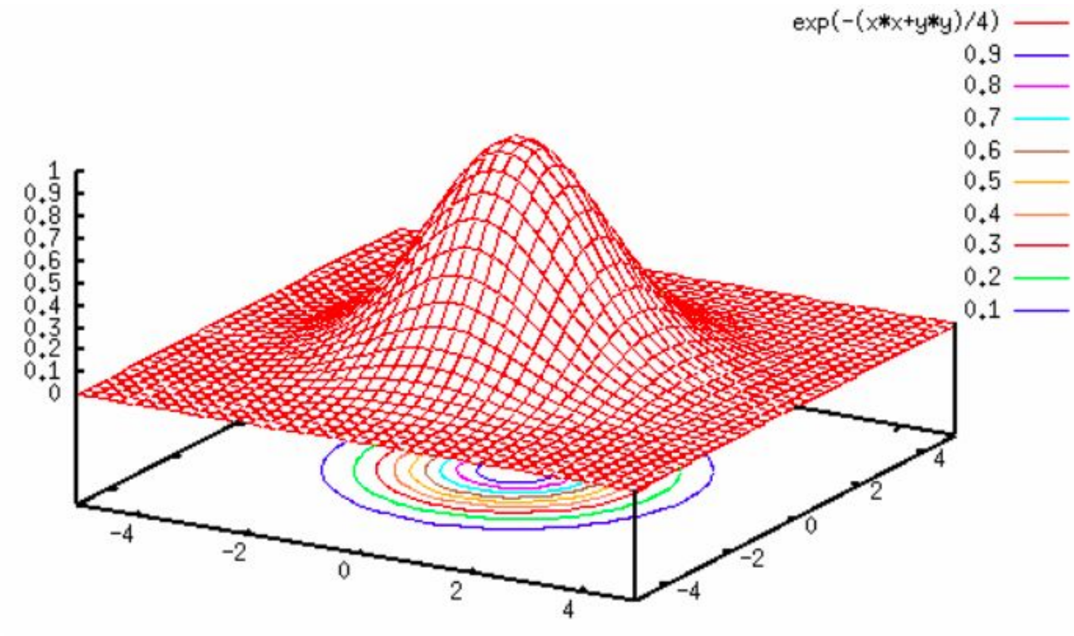
Covariance matrix observations

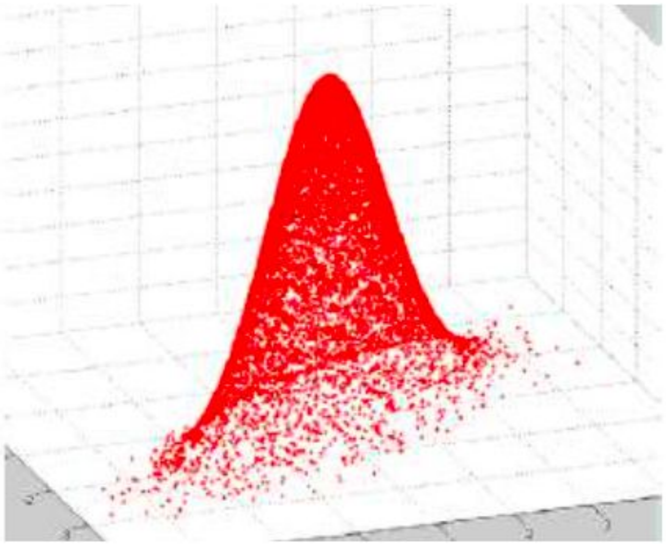
- $\Sigma = \Sigma^T$
- If the covariance matrix is diagonal, all RVs are mutually independent.
- Covariance matrix is positive-semidefinite
  - Every positive definite matrix is invertible

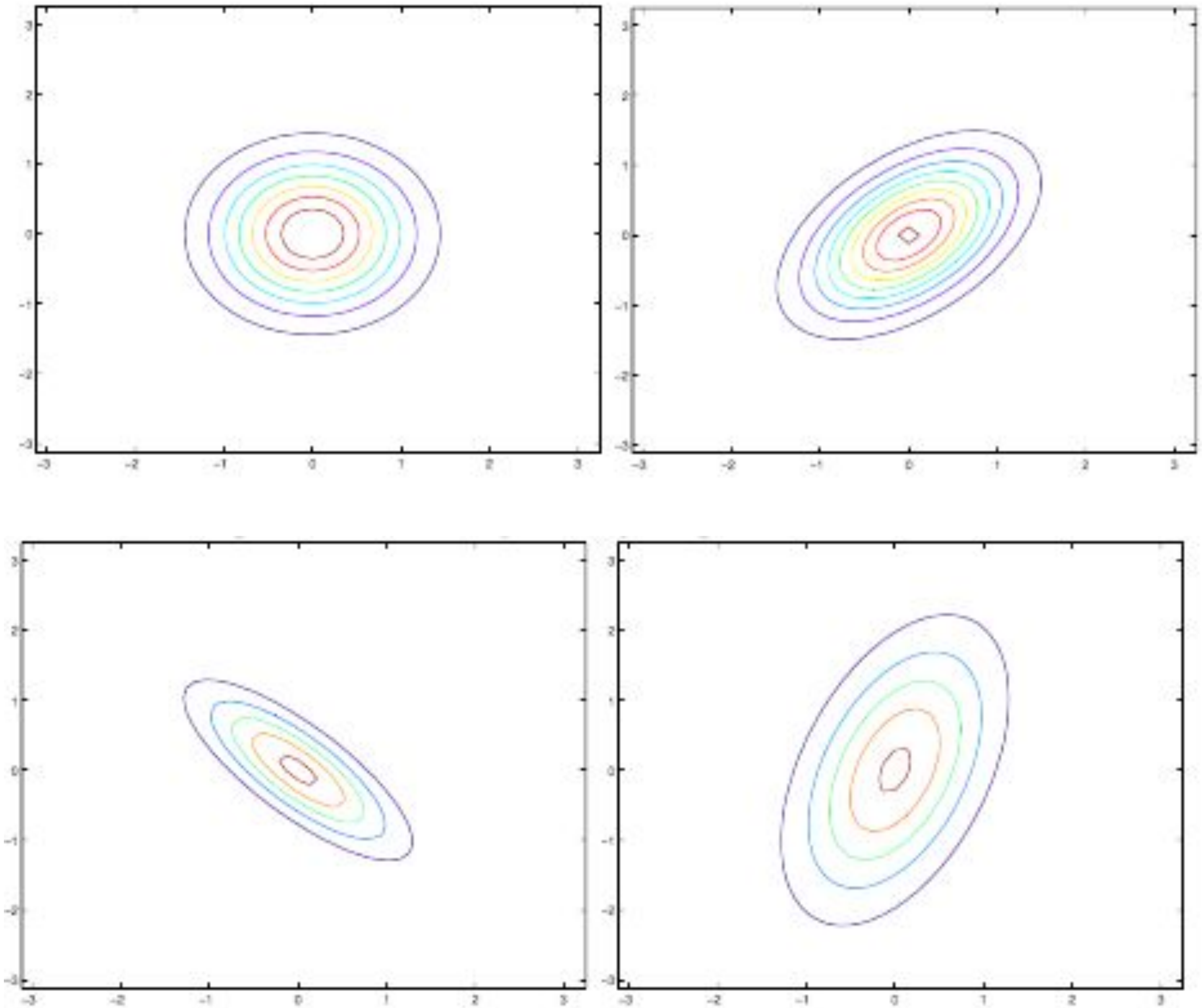
Multivariate Gaussian distribution

- Put $X_1, X_2, X_3 \ldots X_n$ into a vector x

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







1

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

2

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

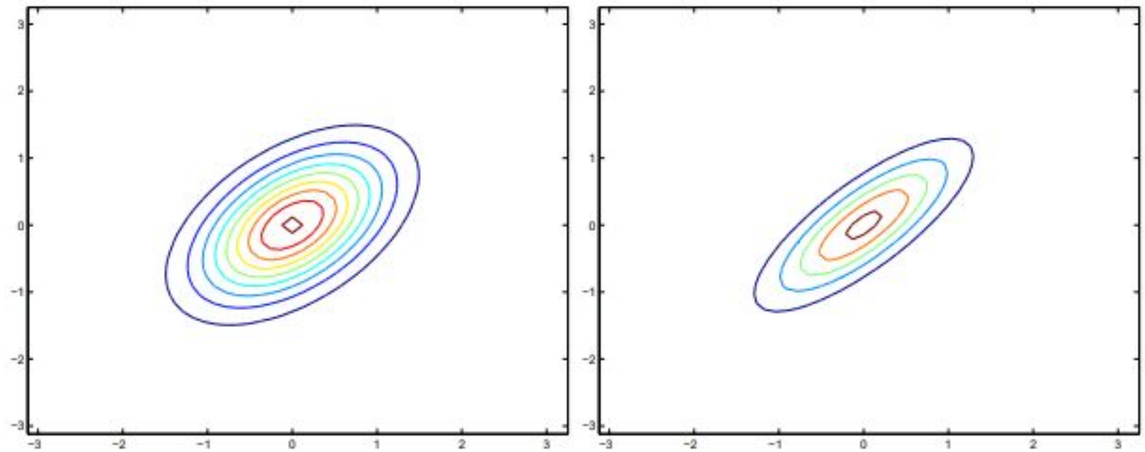
3

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

4

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

http://cs229.stanford.edu/notes/cs229-notes2.pdf



1

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

3

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

http://cs229.stanford.edu/notes/cs229-notes2.pdf

Affine transformation of multivariate Gaussians

- **y = Ax+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

- P(head) = θ, θ = #heads/#tosses
- HHTTH

- $L(\theta) = P(X; \theta) = P(HHTTH; \theta)$
- Maximum Likelihood Estimate (MLE)

Linear Regression Revisit

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

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

- 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

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

error

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

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

Maximizing Likelihood

What is the assumption here?
Is it accurate?

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

From our previous lecture

Min $J(\theta) = \frac{1}{m} \Sigma_{i=1}^{m} (y_i - \theta^T \mathbf{x_i})^2$

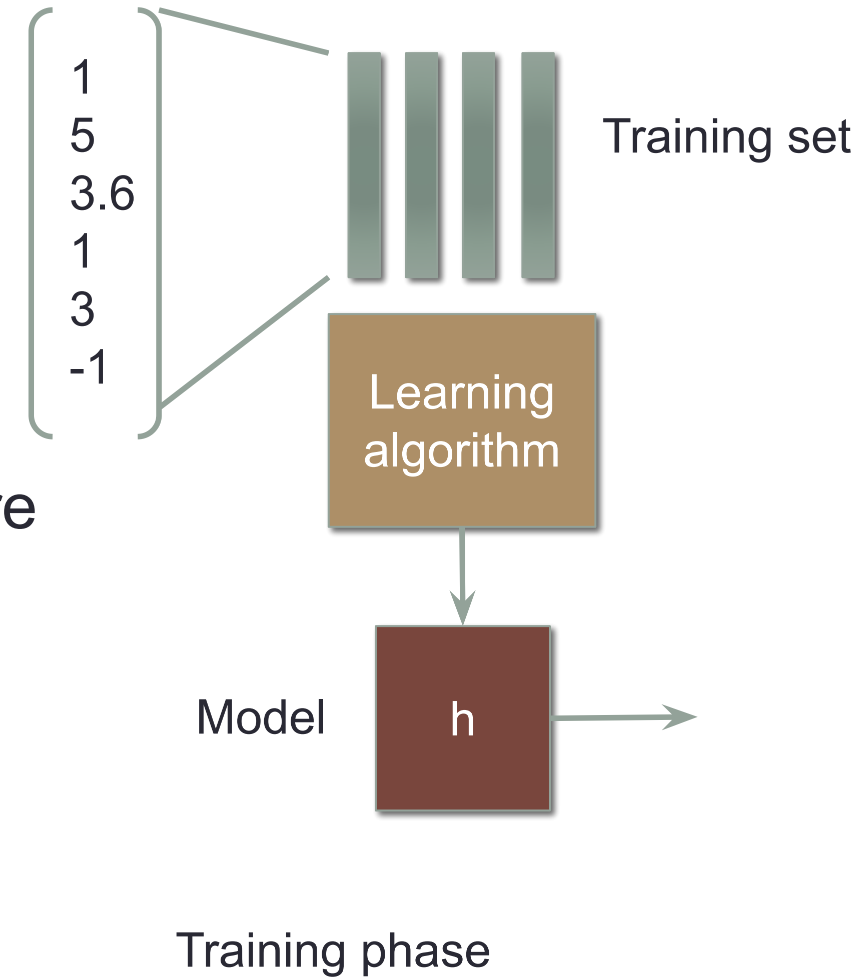
Mean square error solution and MLE solution

- 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.
- Note that MLE derivation assumes that the error is normally distributed! This is a key assumption for linear regression.
  - Error is normally distributed is not that same as y is normally distributed.

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



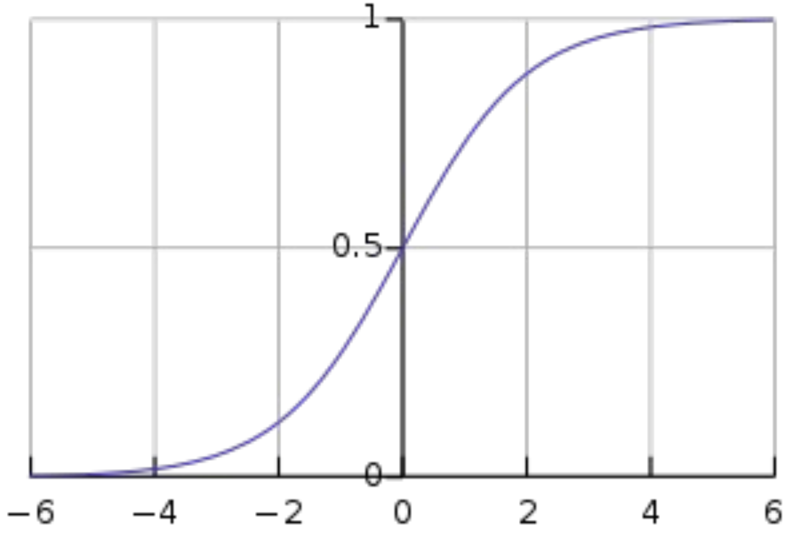
Can we use regression?

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

- 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)



$$f(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$$

https://en.wikipedia.org/wiki/Logistic_function

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

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

Probabilistic view of Logistic Regression

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

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

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

or

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

Maximizing log likelihood

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

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

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

Logistic Regression update rule

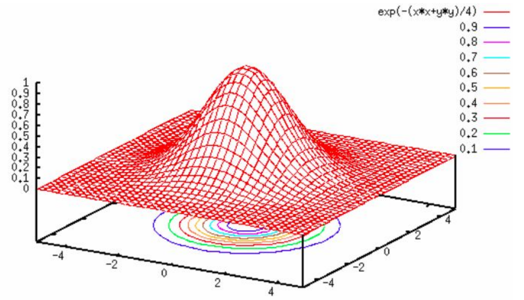
$$\theta_j \Leftarrow \theta_j + r\Sigma_{i=1}^{m}(y_i - h_\theta(x_i))x_i^{(j)}$$

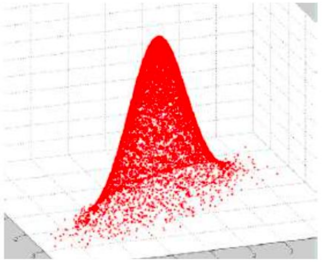
Update rule for linear regression

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

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

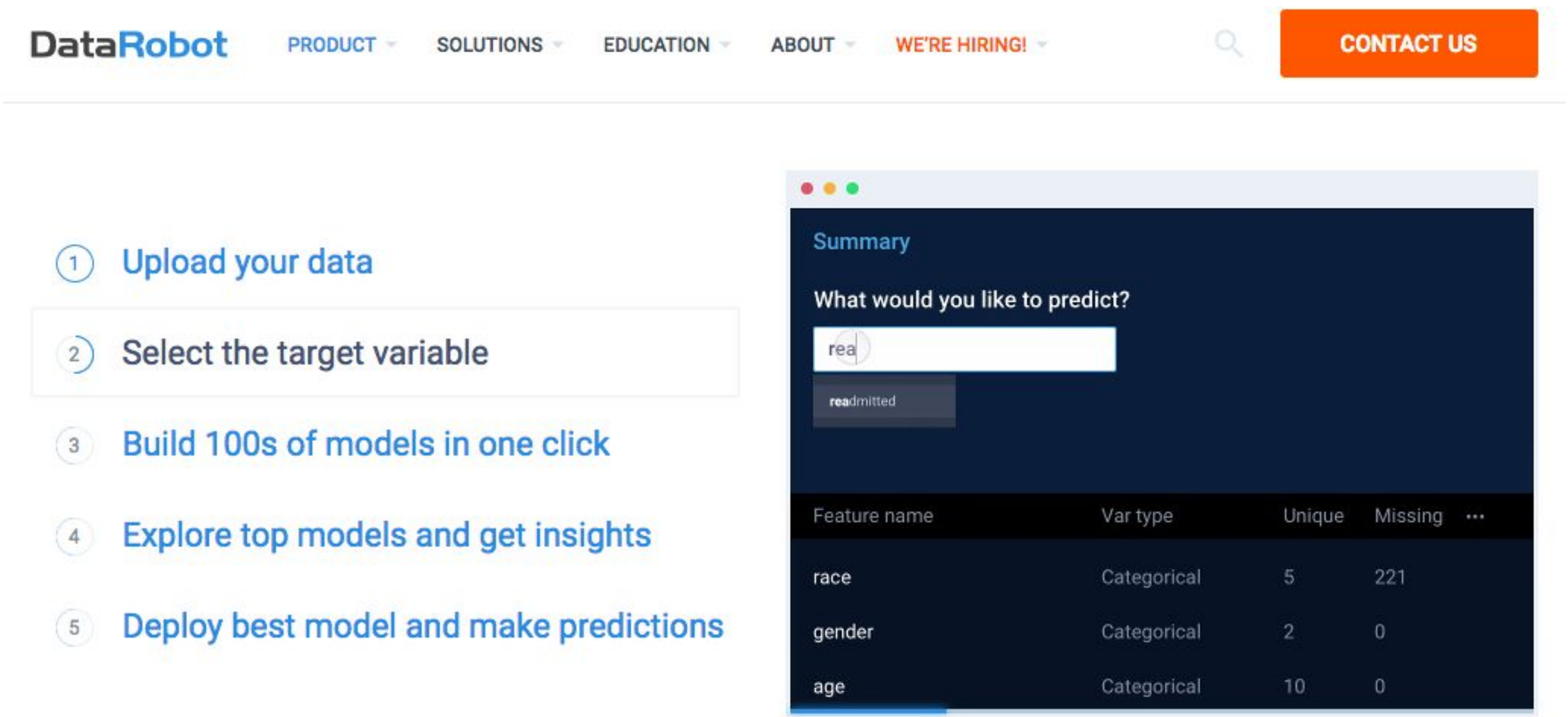
Follow

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

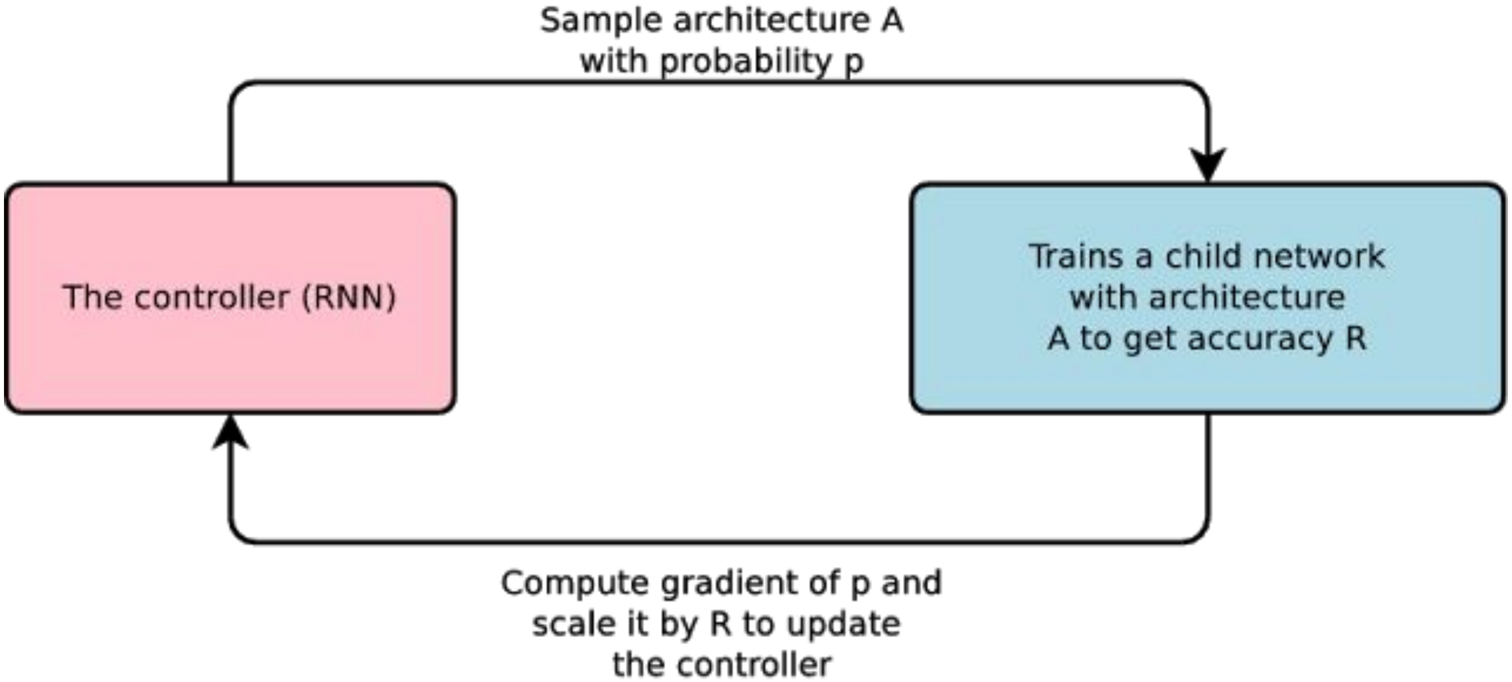
- Machine learning as a tool for non-experts
- Can a non-expert just provide the data and let the machine decides how to proceed



https://www.datarobot.com

Reinforcement Learning for Model Selection

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



https://research.googleblog.com/2017/05/using-machine-learning-to-explore.html