Machine Learning

Guizhou Academy of Sciences
By Shun Zhang

Outline(Content)

- Introduction (Me & Machine learning term)
- Machine Learning Tools & application scenarios
- Linear Modeling Least Square Approach
- Linear Modeling Maximum Likelihood Approach
- <u>Bayesian Approach</u>
- <u>Classification</u>

Formal Self-Introduction

Zhang Shun Neil
张 顺

Seminar Course Manager Daily Affairs/Translator/Guide

Facebook/LinkedIn——> zhangshun025@gmail.com
WhatsApp ——> +8613809411054
Wechat/qq ——> 443156782
Email Address ——> zhangshun025@163.com

What is Machine Learning?

Equipping Computers with Human Like Capabilities

- Endow computers with the abilities to “learn” from “data”.
- Present data from sensors, the internet, experiments.
- Expect computer to make “sensible” decision.
- Traditionally categorized as:

  Supervised learning:classification,regression.

  Unsupervised learning:dimensionality reduction,clustering.

  Reinforcement learning: learning from delayed feedback.

Machine Learning Today

- Early machine learning viewed with skepticism by statisticians.
- Modern machine learning and statistics interact to both communities benefits.

Machine Learning Today

- For the moment the two overlap strongly. But they are not the same field!
- Machine learning also has overlap with Cognitive Science.
- Mathematical formalisms of a problem are helpful, but they can hide facts
- Mathematical foundations are still very important though: they help us understand the capabilities of our algorithms.
- But we mustn’t restrict our ambitions to the limitations of current mathematical formalisms. That is where humans give inspiration.

Statistics and Machine Learning

Statisticians want to turn humans into computers.

Machine learners want to turn computers into humans.

Machine Learning Areas

- Supervised learning:

  tasks of inferring a function from labeled training data

- Unsupervised learning:

  trying to find hidden structure from unlabeled data. No error or reward signal to evaluate a potential solution.

Machine Learning Areas

- Reinforcement learning:

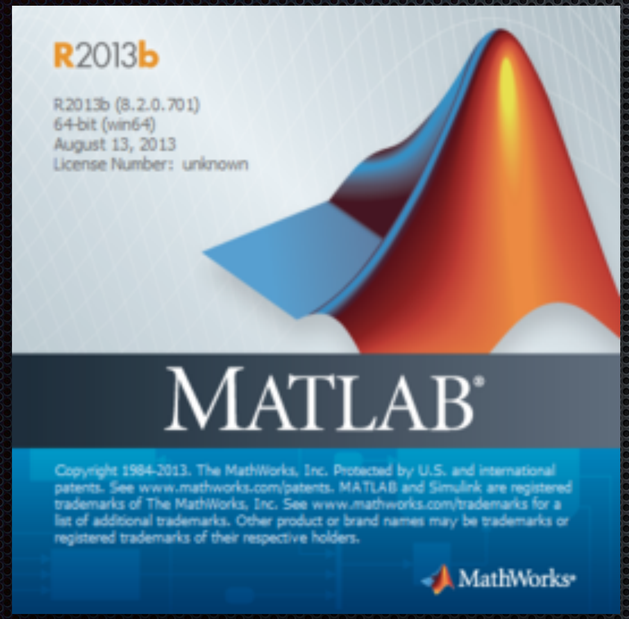
  inspired by behaviorist psychology, concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

Applications of Machine Learning

- Handwriting Recognition: Mobile Phones' handwriting functions http://yann.lecun.com/exdb/lenet/
- Ranking: Results ranking in search engines like google
- Collaborative Filtering: Amazon's item recommendation systems
- Internet Search: Results match your key words by using yahoo
- News Personalization: Online news readers like zite, pulse and clipboard http://www.zite.com/
- Game Play Learning: Popular Game console like Xbox with Kinect

Machine Learning Tools Recommendations

MATLAB



Python

Machine Learning Tools

Python

Necessary Libraries

Scipy, Numpy, Matplotlib…..

Third Party Libraries

scikit-learn, NLTK, Mlpy, Shogun, MDP, PyBrain, BigML, PyML, Milk, PyMVPA, Pattern, pyrallel, Monte, Orange, Theano, Pylearn2……

Machine Learning Tools

Scipy:
a collection of mathematical algorithms and convenience functions built on the Numpy extension of Python.

Integration, Optimization, Interpolation, Fourier Transforms, Signal processing, Linear Algebra, Statictics, Multidimensional image processing……

Special funcs: airy, elliptic, bessel, gamma, beta, hypergeometric……

Machine Learning Tools

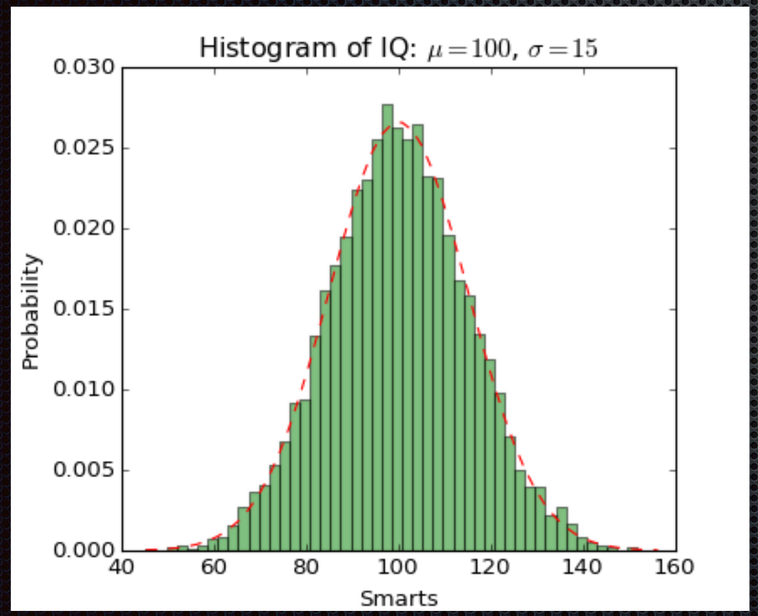
Numpy:
is the fundamental package for scientific computing with Python.

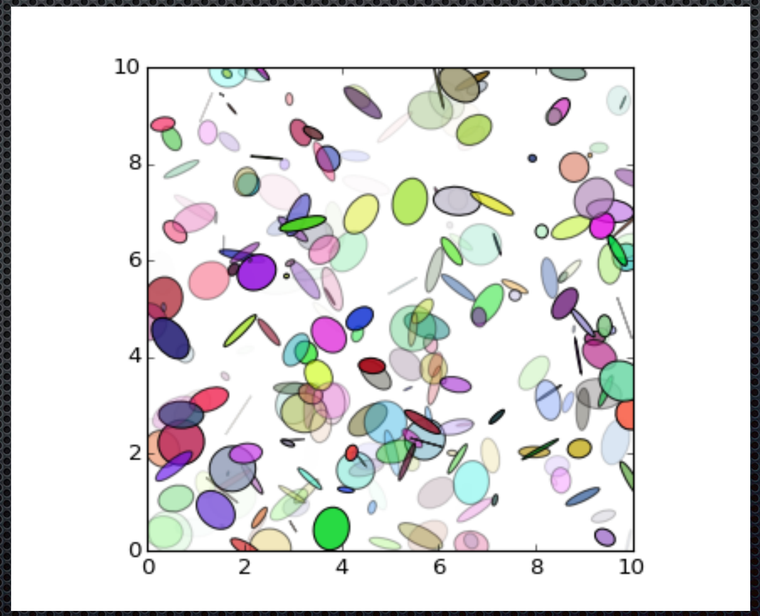
It contains a powerful N-dimensional array object, sophisticated (broadcasting) functions, tools for integrating C/C++ and Fortran code, useful linear algebra, Fourier transform, and random number capabilities.

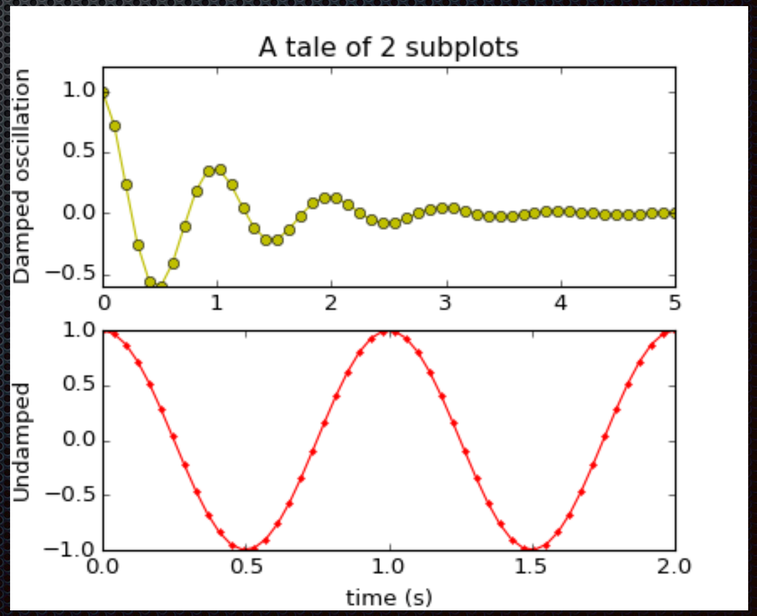
Machine Learning Tools

Matplotlib:

a python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.







Application Scenarios

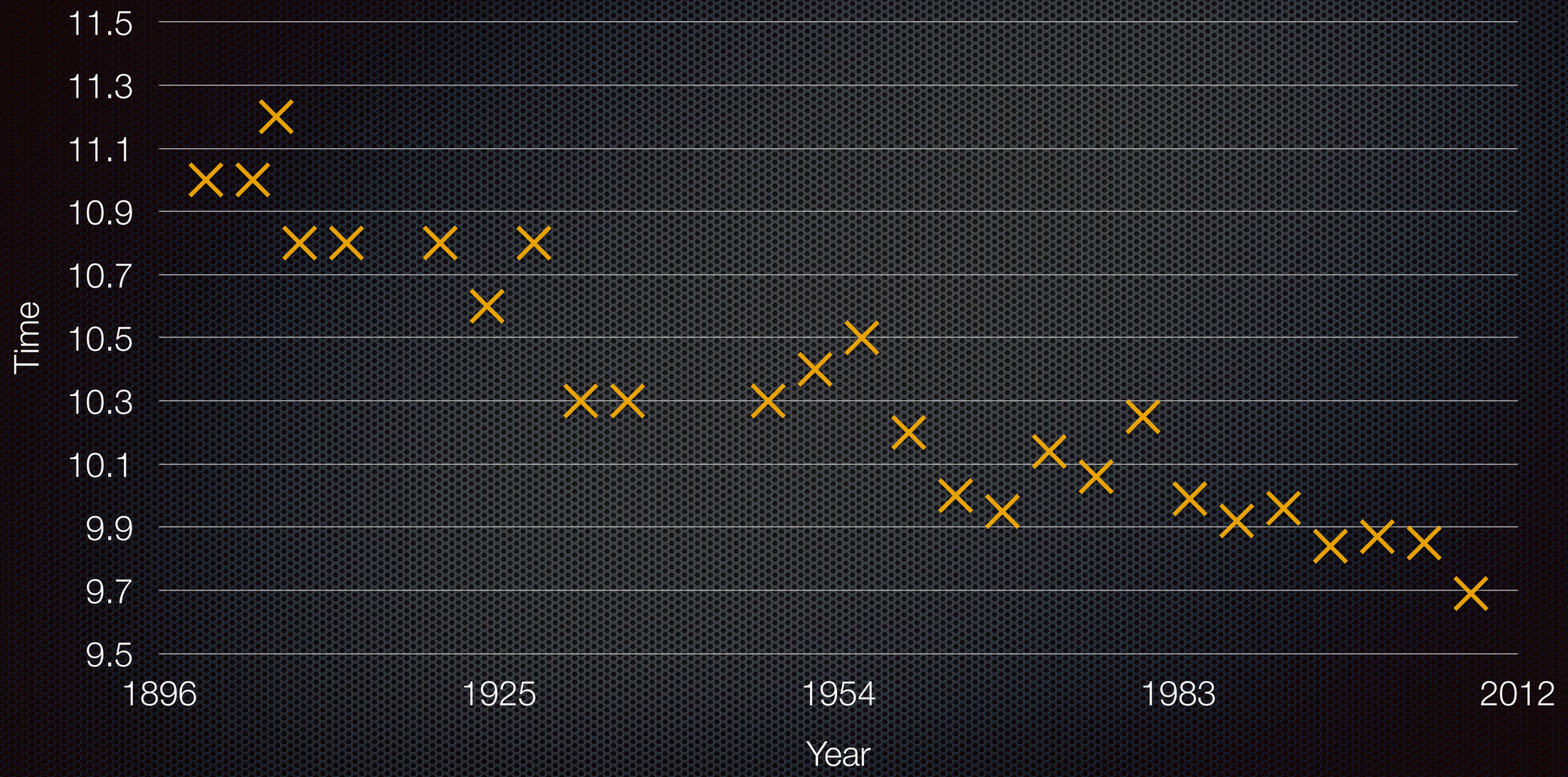
Possible Research Areas

Classification, Natural Language Processing, Data Mining, Artificial Intelligence, Pattern Recognition, Complex Network, Public Opinion Analysis, Big Data Analysis……

Extend Application Areas

Web Service, Database management, Scientific calculator, Embedded programming, Game design, Multimedia processing……

Linear Modeling Least Square Approach



Summer Olympics Men's 100 Game Records since 1896

Linear Modeling Least Square Approach

Defining the model

Input x: the Olympics year
Output t: the winning time in seconds

$$t=f(x)$$
$$t=f(x;a)$$

Linear Modeling Least Square Approach

Modelling assumptions:

The relationship between x and t is linear

The data could be adequately modeled with a straight line.

Linear Modeling Least Square Approach

Linear relationships: $y=mx+c$

Nonlinear relationships: $y=mx^2+c$

$y=\sin(x)$

$$\sqrt{y} = mx + c$$

$y=mx+c^2$ ***$y=x\sin(m)+c$***

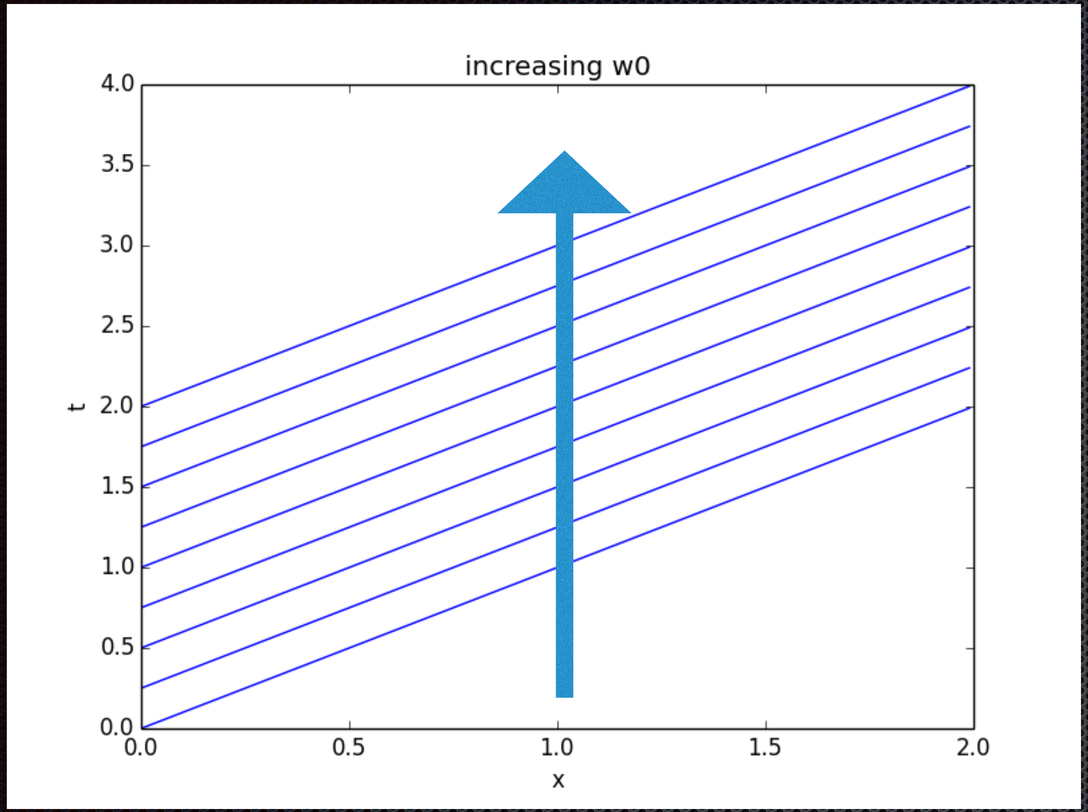
Linear Modeling Least Square Approach

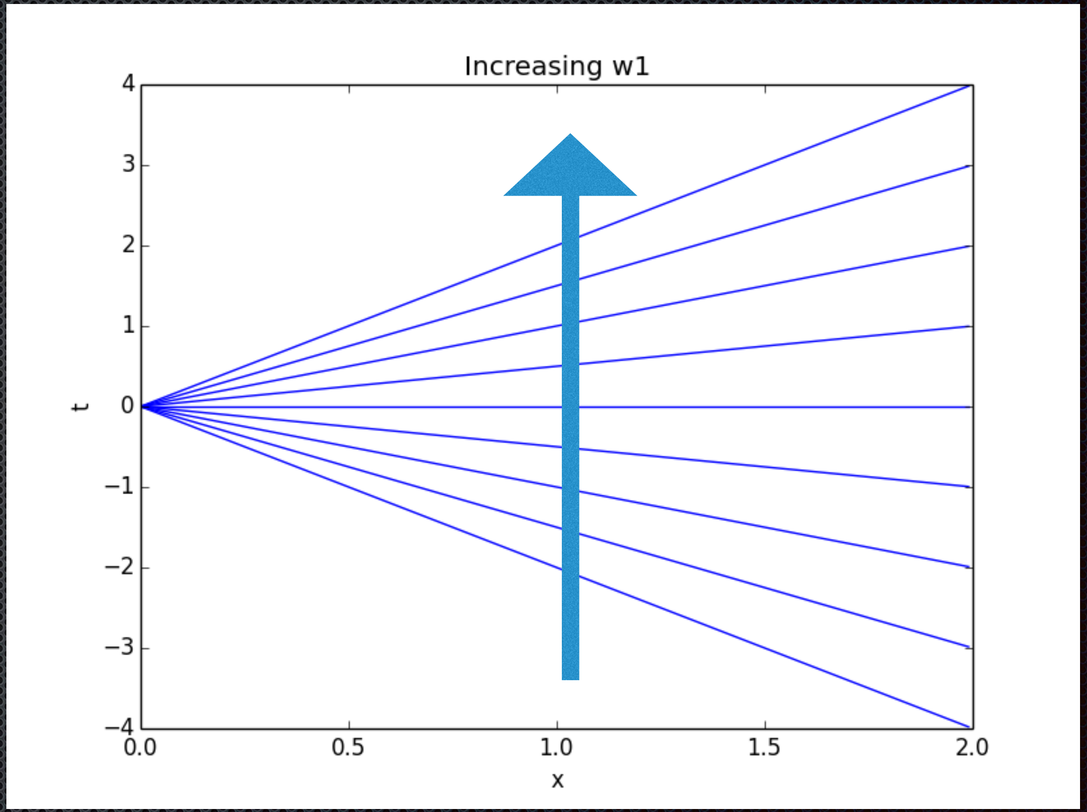
$$t=f(x)=x$$

$$t=f(x;w)=wx$$

$$t=f(x;w_0,w_1)=w_0+w_1x$$

Linear Modeling Least Square Approach





$$t=f(x;w_0,w_1)=w_0+w_1x$$

Linear Modeling Least Square Approach

How to define what a good model is?

the best solution of values of w_0 and w_1 produce a line as close as possible to all data points.

Squared difference between the true winning time and winning time predicted by the model

$$(t_n - f(x_n; w_0, w_1))^2$$

Linear Modeling Least Square Approach

Squared Loss Function:

$$L_n(t_n, f(x_n; w_0, w_1)) = (t_n - f(x_n; w_0, w_1))^2$$

Linear Modeling Least Square Approach

Squared Loss Function:

$$L_n(t_n, f(x_n; w_0, w_1)) = (t_n - f(x_n; w_0, w_1))^2$$

Average Loss Function: $L = \frac{1}{N}\sum_{n=1}^{N} L_n(t_n, f(x_n; w_0, w_1))$

Linear Modeling Least Square Approach

Squared Loss Solution:

n=1,2,……N observations

year x_n time in seconds t_n

$$L = \frac{1}{N}\sum_{n=1}^{N} L_n(t_n, f(x_n; w_0, w_1))$$

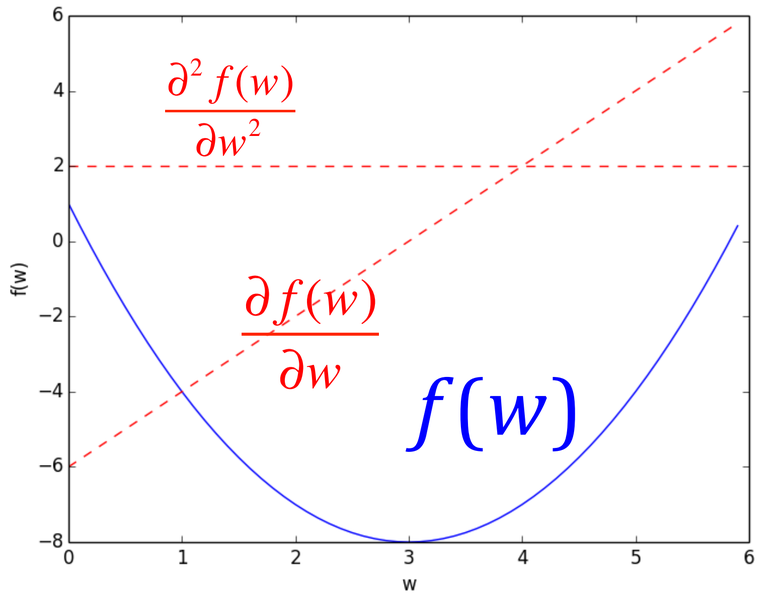
$$= \frac{1}{N}\sum_{n=1}^{N} (w_1^2 x_n^2 + 2w_1 x_n (w_0 - t_n) + w_0^2 - 2w_0 t_n + t_n^2)$$

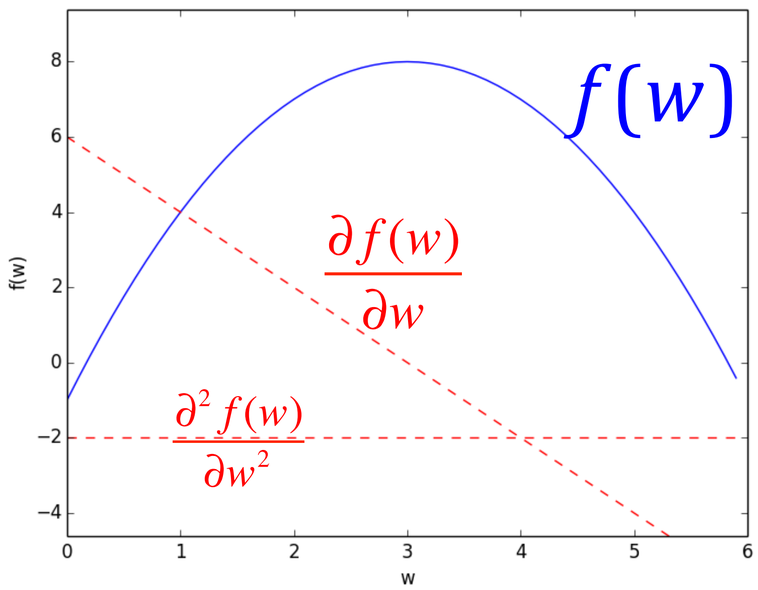
Basic Optimization Overview

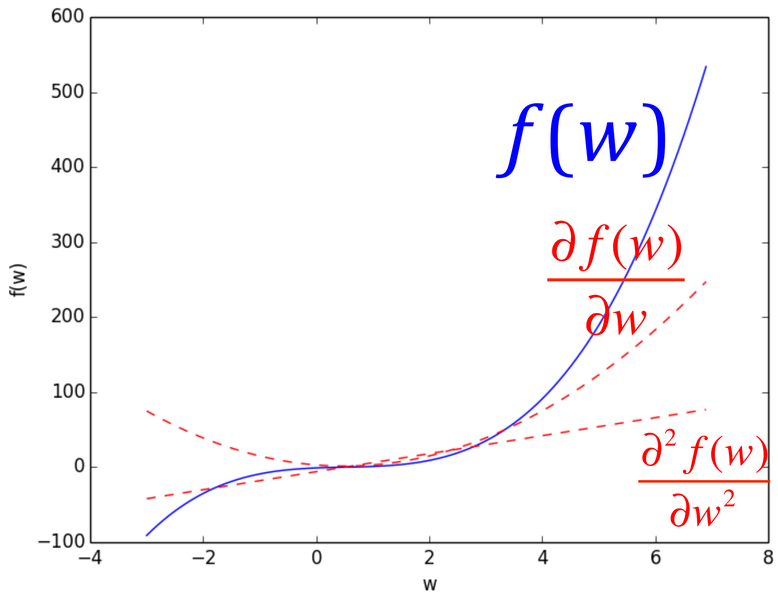
- To find a minimum, we need to find a point where gradient is zero(this is a stationary point).
- If we can show that curvature is positive, this is a minimum.
- Procedure: differentiate the function, find parameters which set derivative to zero.
- This can sometimes be done by a fixed point equation, other times iterative optimization methods are required.

Mathematical Background Knowledge

Optimization——turning points







Linear Modeling Least Square Approach

Squared Loss Solution:

Differentiating the loss function

Partial derivatives with respect to w_1 and w_0

$$\frac{\partial L}{\partial w_1} = 2w_1 \frac{1}{N} \sum_{n=1}^{N} (x_n{}^2) + \frac{2}{N} (\sum_{n=1}^{N} x_n (w_0 - t_n))$$

$$\frac{\partial L}{\partial w_0} = 2w_0 + 2w_1 \frac{1}{N} \sum_{n=1}^{N} (x_n) - \frac{2}{N} (\sum_{n=1}^{N} t_n)$$

Linear Modeling Least Square Approach

Squared Loss Solution:

$$\widehat{w}_0 = \bar{t} - w_1 \bar{x}$$

$$\widehat{w}_1 = \frac{\overline{xt} - \bar{x}\bar{t}}{\overline{x^2} - (\bar{x})^2}$$

Linear Solution

n	x_n	t_n	$x_n t_n$	x_n^2
1	1896	12.00	22752.0	3594800
2	1900	11.00	20900.0	3610000
3	1904	11.00	20944.0	3625200
4	1906	11.20	21347.2	3632800
5	1908	10.80	20606.4	3640500
6	1912	10.80	20649.6	3655700
7	1920	10.80	20736.0	3686400
8	1924	10.60	20394.4	3701800
9	1928	10.80	20822.4	3717200
10	1932	10.30	19899.6	3732600
11	1936	10.30	19940.8	3748100
12	1948	10.30	20064.4	3794100
13	1952	10.40	20300.8	3810300
14	1956	10.50	20538.0	3825900
15	1960	10.20	19992.0	3841600
16	1964	10.00	19640.0	3857500
17	1968	9.95	19581.6	3873000
18	1972	10.14	19996.1	3888800
19	1976	10.06	19878.6	3904600
20	1980	10.25	20295.0	3920400
21	1984	9.99	19820.2	3936300
22	1988	9.92	19721.0	3952100
23	1992	9.96	19840.3	3968100
24	1996	9.84	19640.6	3984000
25	2000	9.87	19740.0	4000000
26	2004	9.85	19739.4	4016000
27	2008	9.69	19457.5	4032100
AVERAGE	1952.37	10.39	20268.1	3813000

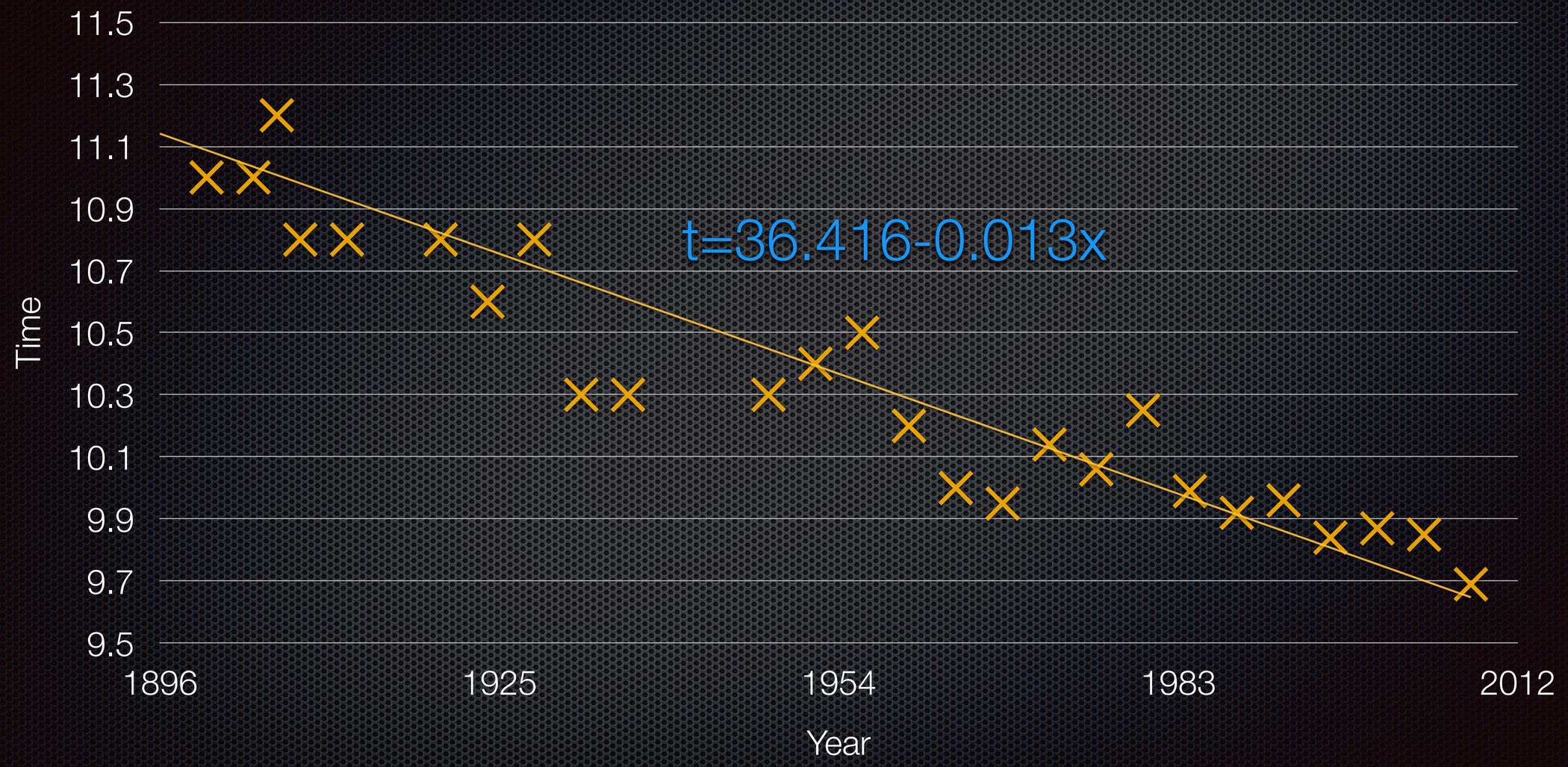
Linear Solution

In this example, we get:

$$w_1 = \frac{20268.1 - 1952.37 \times 10.39}{3.8130 \times 10^6 - 1952.37 \times 1952.37} = -0.013$$

$$w_0 = 10.39 - (-0.0133) \times 1952.37 = 36.416$$

Linear Solution



Summer Olympics Men's 100 Game Records since 1896

Linear Solution

Making predictions:

$t_{2012}=f(2012;w_0,w_1)=9.595;$
$t_{2016}=f(2016;w_0,w_1)=9.541;$

2012 in London Usain Bolt 9.63 broke the Olympics record!

Linear Modeling Least Square Approach

A solution that is determined by a set of several attributes

$$t=f(x,s_1,\ldots\ldots s_7;w_0,\ldots\ldots w_8)=w_0+w_1x$$
$$+w_2s_1+w_3s_2+w_4s_3+w_5s_4+w_6s_5+w_7s_6+w_8s_7$$

Vector and Matrix notation is needed!

Algebra Revision

Scalars, Vectors and Matrices

standard convention to represent scalar value x

lower letters by bold to represent vector value **x** $\bar{x}$

uppercase letters by bold to represent matrices **X**

Algebra Revision

Vectors / Matrices transpose, dimensions and indexing

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$x = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \qquad x^T = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \end{bmatrix}$$

$$A^T = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}$$

$$x = \left[year, s_1, s_2, \ldots\ldots, s_8 \right]^T$$

Algebra Revision

Matrix multiplication:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \qquad B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}$$

N x M P x Q

$$C_{ij} = \sum_k B_{ik} A_{kj}$$

Algebra Revision

Inner product between two column vectors.

$$z = x^T y = \sum_k x_k y_k$$

the result is a scalar!

Algebra Revision

Transpose of a product:

$$(Xw)^T = w^T X^T$$

$$(ABCD)^T = ((AB)^T (CD)^T) = (CD)^T (AB)^T = D^T C^T B^T A^T$$

Algebra Revision

Identity matrix:

$$I_1 = 1, I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{y}^T I_D = \mathbf{y}, \quad I\mathbf{y} = \mathbf{y}$$

$$\mathbf{A} I_M = \mathbf{A}, \quad I_N \mathbf{A} = \mathbf{A}$$

Algebra Revision

Matrix inverse:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$A = \begin{bmatrix} a1 & 0 & 0 & 0 \\ 0 & a2 & 0 & 0 \\ 0 & 0 & a3 & 0 \\ 0 & 0 & 0 & a4 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} a_1^{-1} & 0 & 0 & 0 \\ 0 & a_2^{-1} & 0 & 0 \\ 0 & 0 & a_3^{-1} & 0 \\ 0 & 0 & 0 & a_4^{-1} \end{bmatrix}$$

$$I^{-1} = I$$

Linear Modeling Least Square Approach

Adapt the Matrix expression to our original model:

$$w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad x_n = \begin{bmatrix} 1 \\ x_n \end{bmatrix}$$

$$f(x_n; w_0, w_1) = w^T x_n = w_0 + w_1 x_n$$

$$L = \frac{1}{N} \sum_{n=1}^{N} (t_n - w^T x_n)^2 = \frac{1}{N} (t - Xw)^T (t - Xw)$$

Linear Modeling Least Square Approach

Adapt the Matrix expression to our original model:

$$L = \frac{1}{N}\sum_{n=1}^{N}(t_n - w^T x_n)^2 = \frac{1}{N}(t - Xw)^T(t - Xw)$$

Do they equal to each other?

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \cdot \\ \cdot \\ \cdot \\ X_N^T \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_N \end{bmatrix} \qquad t = \begin{bmatrix} t_1 \\ t_2 \\ \cdot \\ \cdot \\ \cdot \\ t_N \end{bmatrix}$$

Linear Modeling Least Square Approach

Adapt the Matrix expression to our original model:

$$Xw = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \times \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} w_0 + w_1 x_1 \\ w_0 + w_1 x_2 \\ \vdots \\ w_0 + w_1 x_N \end{bmatrix}$$

Linear Modeling Least Square Approach

Adapt the Matrix expression to our original model:

$$t - Xw = \begin{bmatrix} t_1 - w_0 - w_1 x_1 \\ t_2 - w_0 - w_1 x_2 \\ \vdots \\ t_N - w_0 - w_1 x_N \end{bmatrix}$$

Linear Modeling Least Square Approach

Therefore:

$$(t - Xw)^T (t - Xw) = (w_0 + w_1 x_1)^2 + (w_0 + w_1 x_2 - t_2)^2 + \ldots$$

$$+(w_0 + w_1 x_N - t_N)^2$$

$$= \sum_{n=1}^{N} (w_0 + w_1 x_n - t_n)^2$$

$$= \sum_{n=1}^{N} (t_n - f(x_n; w_0, w_1))^2$$

Linear Modeling Least Square Approach

Therefore:

$$L = \frac{1}{N}\sum_{n=1}^{N}(t_n - w^T x_n)^2 = \frac{1}{N}(t - Xw)^T (t - Xw)$$

Proved!

Linear Modeling Least Square Approach

Continue to figure out the equation:

$$L = \frac{1}{N}(t - Xw)^T (t - Xw)$$

$$= \frac{1}{N}(Xw)^T Xw - \frac{1}{N}(Xw)^T t - \frac{1}{N} t^T Xw + \frac{1}{N} t^T t$$

$$= \frac{1}{N} w^T X^T Xw - \frac{2}{N} w^T X^T t + \frac{1}{N} t^T t$$

Linear Modeling Least Square Approach

Differentiating loss in vector/matrix form:

Some useful identities for differentiating a vector:

$f(w)$	$\frac{\partial f}{\partial w}$
$w^T X$	X
$\mathrm{X}^{\mathrm{T}}\mathrm{w}$	X
$\mathrm{w}^{\mathrm{T}}\mathrm{w}$	$2\mathrm{w}$
$\mathrm{w}^{\mathrm{T}}\mathrm{Cw}$	$2\mathrm{Cw}$

Linear Modeling Least Square Approach

Differentiating loss in vector/matrix form:

$$L = \frac{1}{N}(w^T X^T X w - 2wX^T t + t^T t)$$

$$\frac{\partial L}{\partial w} = \frac{2}{N} X^T X w - \frac{2}{N} X^T t = 0$$

$$X^T X w = X^T t$$

Linear Modeling Least Square Approach

Differentiating loss in vector/matrix form:

$$\hat{w} = (X^T X)^{-1} X^T t$$

Linear Modeling Least Square Approach

Nonlinear response from a linear model:

Linear Model: $f(x;w) = w_0 + w_1 x$

Quadratic Model: $f(x;w) = w^T x = w_0 + w_1 x + w_2 x^2$

Polynomial Model: $f(x;w) = \sum_{k=0}^{K} w_k x^k$

Linear Modeling Least Square Approach

Nonlinear response from a linear model:

Demo: Basis of the the model from 1 to 26

Linear_Nonlinear function.py

Linear Modeling Least Square Approach

Nonlinear response from a linear model:

Increasing the polynomial order will always result in a model that gets closer to the training data.

Predictions may not look so sensible by the increasing of order

Linear Modeling Least Square Approach

Other functions in the model:

$$X = \begin{bmatrix} h_1(x_1) & h_2(x_1) & . & . & . & h_K(x_1) \\ h_1(x_2) & h_2(x_2) & . & . & . & h_K(x_2) \\ . & . & & & & . \\ . & . & & & & . \\ . & . & . & . & . & . \\ h_1(x_N) & h_2(x_N) & . & . & . & h_K(x_N) \end{bmatrix}$$

Linear Modeling Least Square Approach

Other functions in the model(Example):

$$h_1(x) = 1$$

$$h_2(x) = x$$

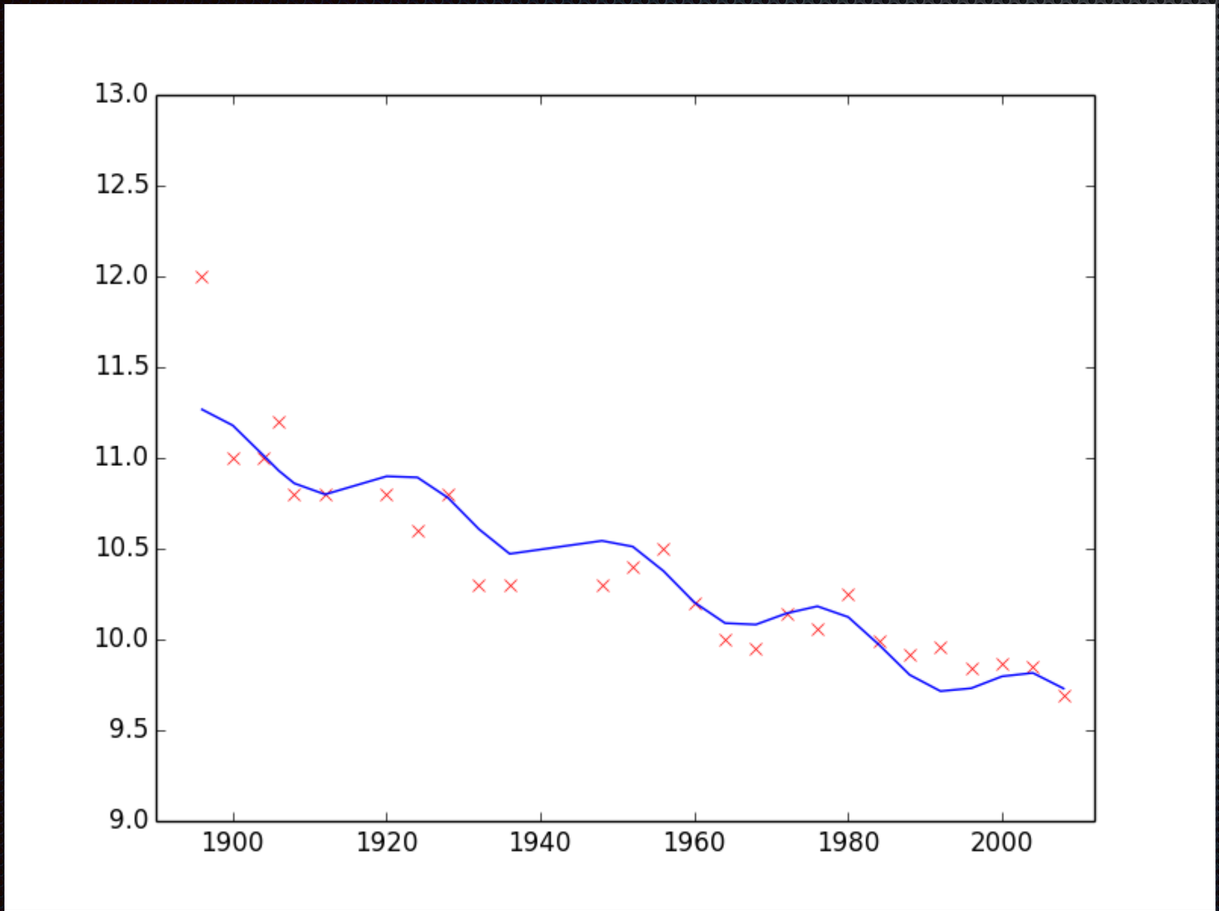
$$h_3(x) = \sin(\frac{x-a}{b})$$

$$f(x;w) = w_0 + w_1 x + w_2 \sin(\frac{x-a}{b})$$

(a=2660, b=4.3)

Linear Modeling Least Square Approach

Other functions in the model(Example):



$$f(x;w) = w_0 + w_1 x + w_2 \sin(\frac{x-a}{b})$$

(a=2660, b=4.3)

Generalization VS Over-fitting

Generalization:
The model that is able to make the predictions. It can generalize beyond the examples we have for training.

Over-fitting:
The model pay too much attention to the training data and has poor ability to generalize new data.

Validation Data

Use a second dataset provided separately or created by removing some from the original training set to validate the model.

Demo explain!

Validation error for N basis function.py

Validation Data

K folds Cross Validation

K cross validation.py

Leave-One-Out Cross Validation(LOOCV)

Leave One Out Cross Validation.py

Linear Modelling: A Maximum Likelihood Approach

Two Simultaneous Equations,What cause a noise?

A system of two simultaneous equations with two unknowns.

$$t_1 = mx_1 + c$$

$$t_2 = mx_2 + c$$

Introducting Noise to our model

Linear Modelling: A Maximum Likelihood Approach

Two Simultaneous Equations,What cause a noise?

A system of two simultaneous equations with two unknowns.

$$t_1 - t_2 = m(x_1 - x_2)$$

Linear Modelling: A Maximum Likelihood Approach

Two Simultaneous Equations,What cause a noise?

A system of two simultaneous equations with two unknowns.

$$m = \frac{t_1 - t_2}{(x_1 - x_2)}$$

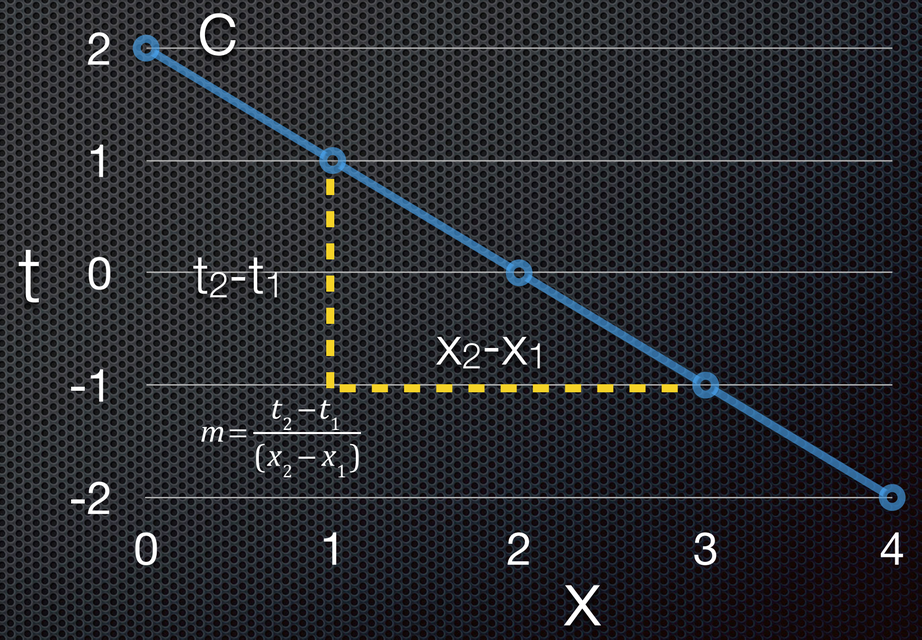
Linear Modelling: A Maximum Likelihood Approach

Two Simultaneous Equations,What cause a noise?

A system of two simultaneous equations with two unknowns.

$$m = \frac{t_2 - t_1}{(x_2 - x_1)}$$

$$c = t_1 - m x_1$$



Linear Modelling: A Maximum Likelihood Approach

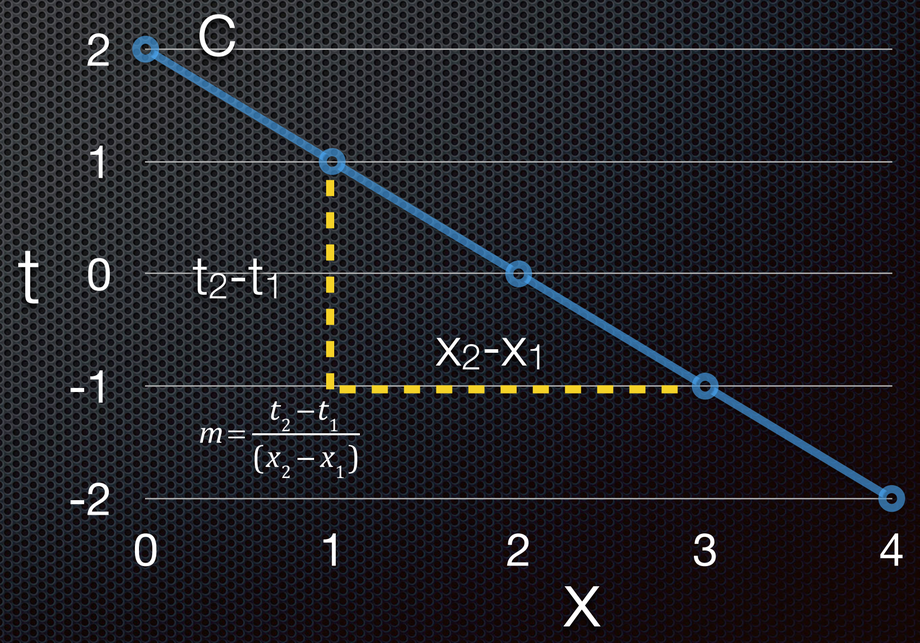
Two Simultaneous Equations,What cause a noise?

How do we deal with three simultaneous equations with only two unknowns?

$$t_1 = mx_1 + c$$

$$t_2 = mx_2 + c$$

$$t_3 = mx_3 + c$$



Linear Modelling: A Maximum Likelihood Approach

- With two unknowns and two observations:

$$t_1 = mx_1 + c$$

$$t_2 = mx_2 + c$$

- Additional observation leads to overdetermined system

$$t_3 = mx_3 + c$$

- This problem is solved through a noise model

$$t_1 = mx_1 + c$$

$$t_2 = mx_2 + c$$

$$t_3 = mx_3 + c$$

Overdetermined System!

Noise Models

- We aren’t modeling entire system.
- Noise model gives mismatch between model and data.
- Gaussian model justified by appeal to central limit theorem.
- Other models also possible (Student-t for heavy tails).
- Maximum likelihood with Gaussian noise leads to least squares.

Machine Learning and Probability

- The world is an uncertain place.

  Epistemic uncertainty:uncertainty arising through lack of knowledge.

  Aleatoric uncertainty: uncertainty arising through an underlying stochastic system.

Probability:A Framework to Characterize Uncertainty

- We need a framework to characterize the uncertainty.
- During the lecture we make use of probability theory to characterize uncertainty.

Probability Review

- We are interested in trials which result in two random variables, X and Y, each of which has an 'outcome' denoted by x or y.
- We summarize the notation and terminology for these distributions in the following table.

Terminology	Notation	Description
Joint probability	$P(X=x,Y=y)$	'The probability that X=x and Y=y'
Marginal Probability	$P(X=x)$	'The probability that X=x regardless of Y'
Conditional Probability	$P(X=x\|Y=y)$	'The probability that X=x given that Y=y'

Table:The different basic probability distributions

A Pictorial Definition of Probability

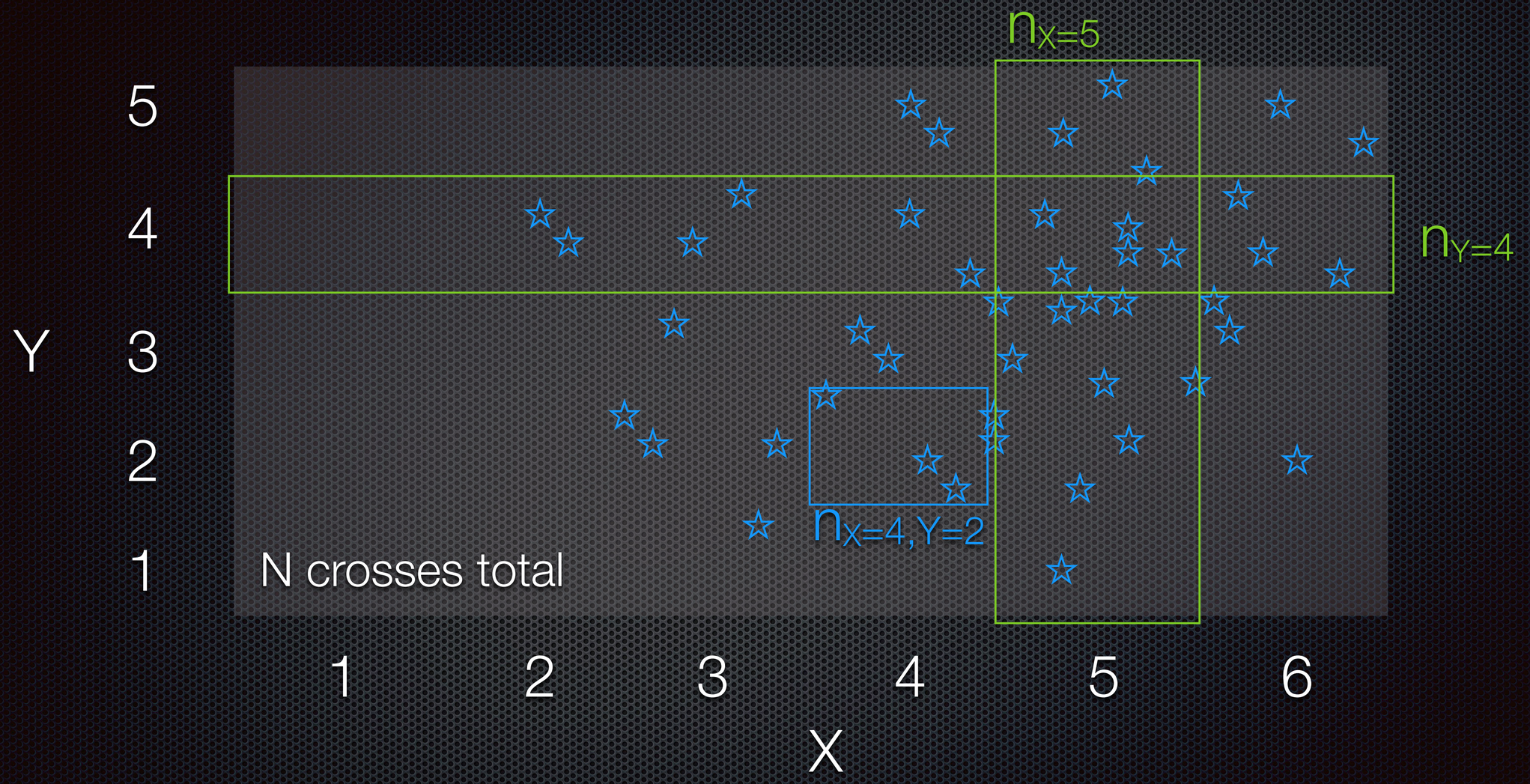


Figure:Representation of joint and conditional probabilities

Different Distributions

Terminology	Definition	Notation
Joint probability	$\lim_{N\to\infty} \frac{n_{X=3,Y=4}}{N}$	P(X=3,Y=4)
Marginal Probability	$\lim_{N\to\infty} \frac{n_{X=5}}{N}$	P(X=5)
Conditional Probability	$\lim_{N\to\infty} \frac{n_{X=3,Y=4}}{n_{Y=4}}$	P(X=3|Y=4)

Table:Definition of probability distributions

Notational Details

- Typically we should write out P(X=x,Y=y)
- In practice, we often use P(X,Y)
- This looks very much like we might write a multivariate function,e.g. $f(x,y) = \frac{x}{y}$
    - For a multivariate function though, $f(x,y) \neq f(y,x)$
    - However P(x,y)= P(y,x) because

      P(X=x,Y=y) = P(Y=y,X=x).
- We now quickly review the ‘rules of probability’

Normalization

- All distributions are normalized. This is clear from the fact that $\sum_x n_x = N$, which gives
- $$\sum_x P(x) = \frac{\sum_x n_x}{N} = \frac{N}{N} = 1$$
- A similar result can be derived for the marginal and conditional distributions

The Sum Rule

Ignoring the limit in our definitions:

The marginal probability P(y) is $\frac{n_y}{N}$ (ignoring the limit)

The joint distribution P(x,y) is $\frac{n_{x,y}}{N}$

$$n_y = \sum_x n_{x,y} \qquad \text{so} \qquad \frac{n_y}{N} = \sum_x \frac{n_{x,y}}{N}$$

in other words

$$P(y) = \sum_x P(x,y)$$

This is known as the sum rule of probability

The Product Rule

- P(x|y) is $$\frac{n_{x,y}}{x_y}$$
- P(x,y) is $$\frac{n_{x,y}}{N} = \frac{n_{x,y}}{x_y}\frac{n_y}{N}$$

or in other words

$$P(x,y) = P(x \mid y)P(y)$$

This is known as the product rule of probability

The Bayes' Rule

From the product rule,

$$P(y,x) = P(x,y) = P(x|y)\,P(y),$$

So

$$P(y|x)\,P(x) = P(x|y)\,P(y)$$

Which leads to Bayes' rule,

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

Distribution Representation

We can represent probabilities as tables

y	0	1	2
P(y)	0.2	0.5	0.3

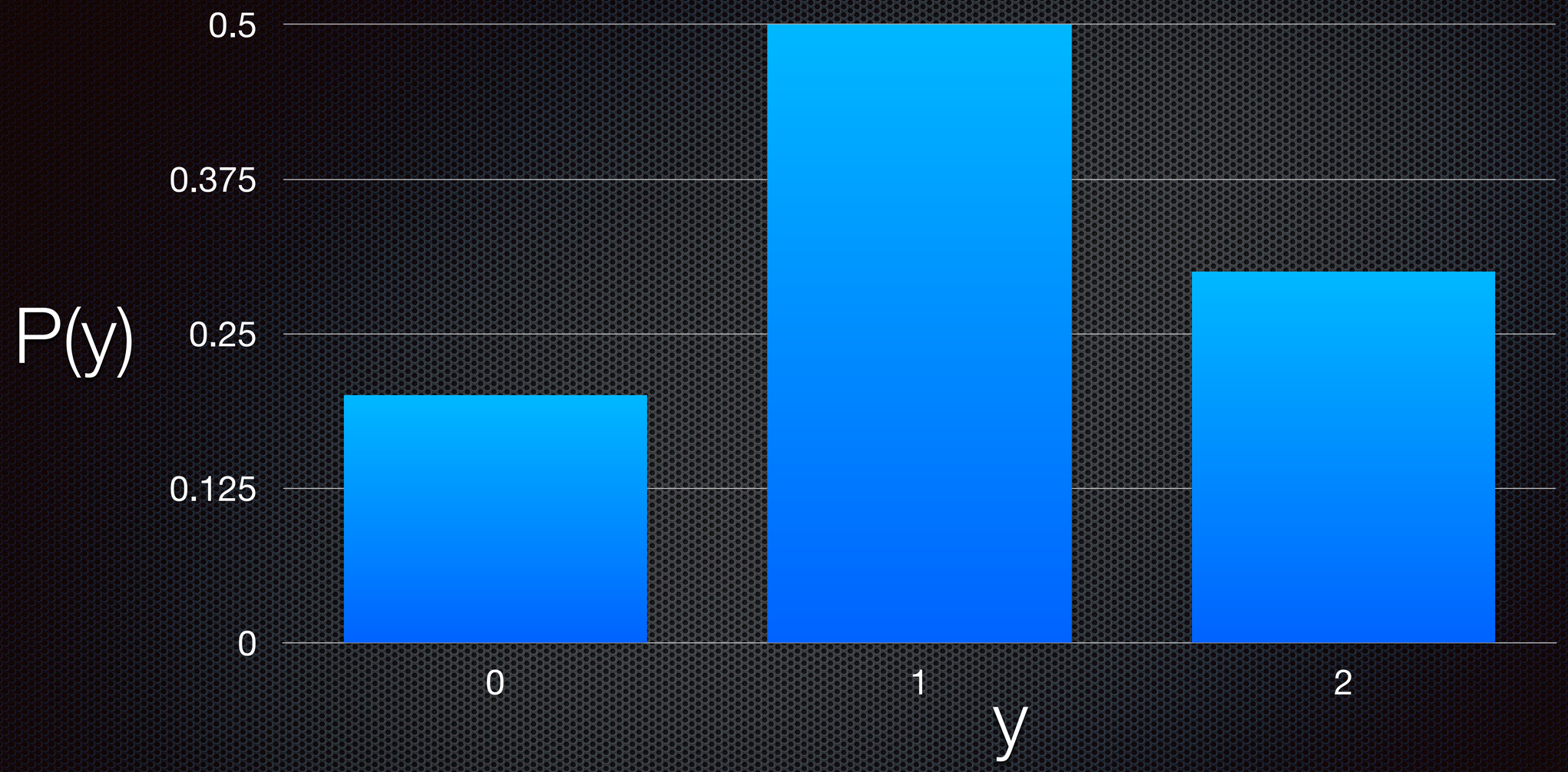


Figure: Histogram representation of the simple distribution

Expectations of Distributions

- Writing down the entire distributions is tedious.
- Can summarize through expectations.

$$< f(y) >_{P(y)} = \sum_{y} f(y) P(y)$$

$$< y >_{P(y)} = 0.2 \times 0 + 0.5 \times 1 + 0.3 \times 2 = 1.1$$

- Consider:

y	0	1	2
P(y)	0.2	0.5	0.3

- We have

$$< y >_{P(y)} = 0.2 \times 0 + 0.5 \times 1 + 0.3 \times 2 = 1.1$$

- This is the first moment or mean of the distribution

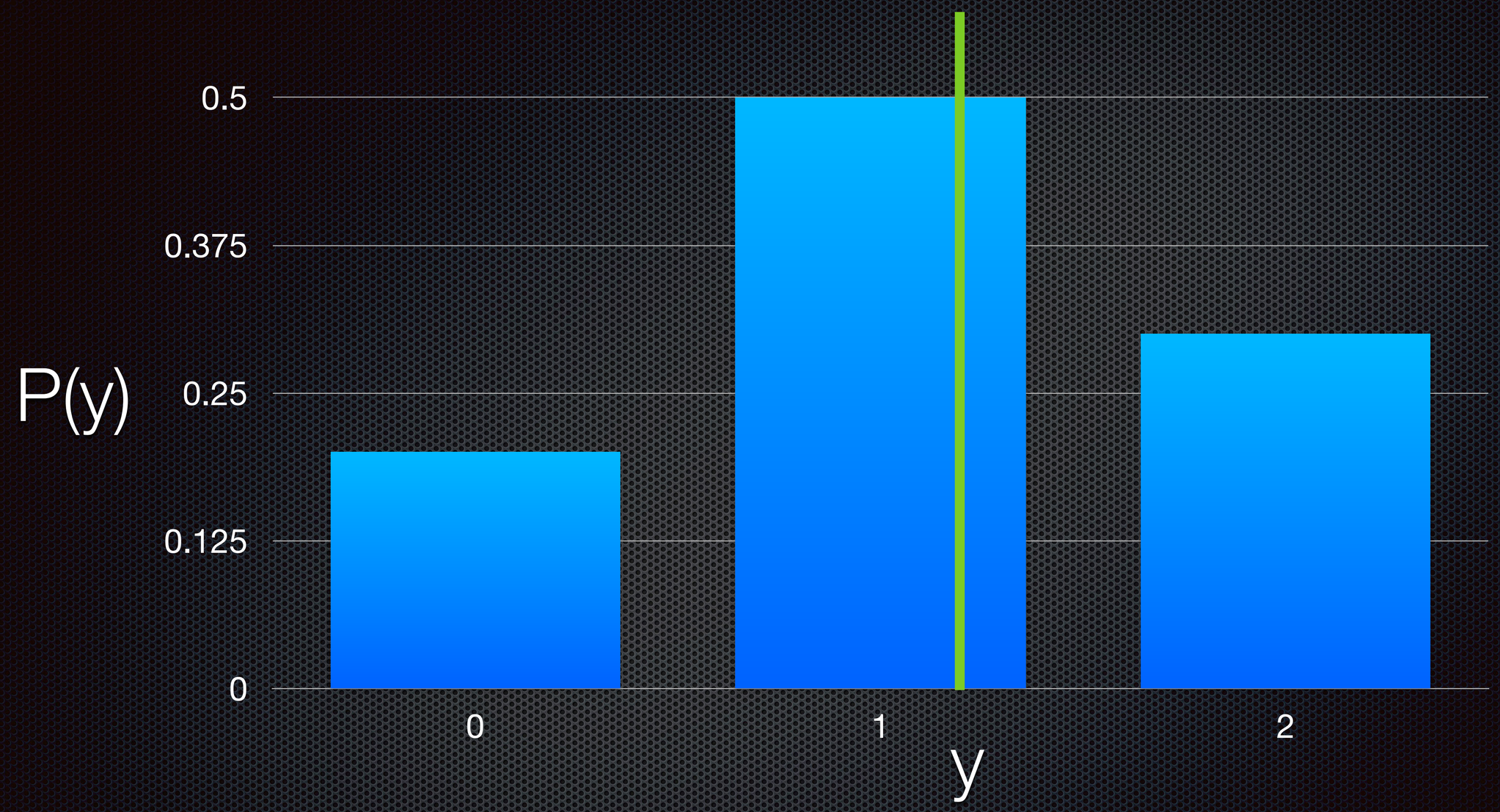


Figure: Histogram representation of the simple distribution including the expectation of y (green line), the mean of the distribution

Variance and Standard Deviation

- Mean gives us the center of the distribution
- Consider:

y	0	1	2
y^2	0	1	4
P(y)	0.2	0.5	0.3

- Second moment is

$$< y^2 >_{P(y)} = 0.2 \times 0 + 0.5 \times 1 + 0.3 \times 4 = 1.7$$

- Variance is

$$< y^2 > - < y >^2 = 1.7 - 1.1 \times 1.1 = 0.49$$

- Standard deviation is square root of variance
- Standard deviation gives us the ‘width’ of the distribution

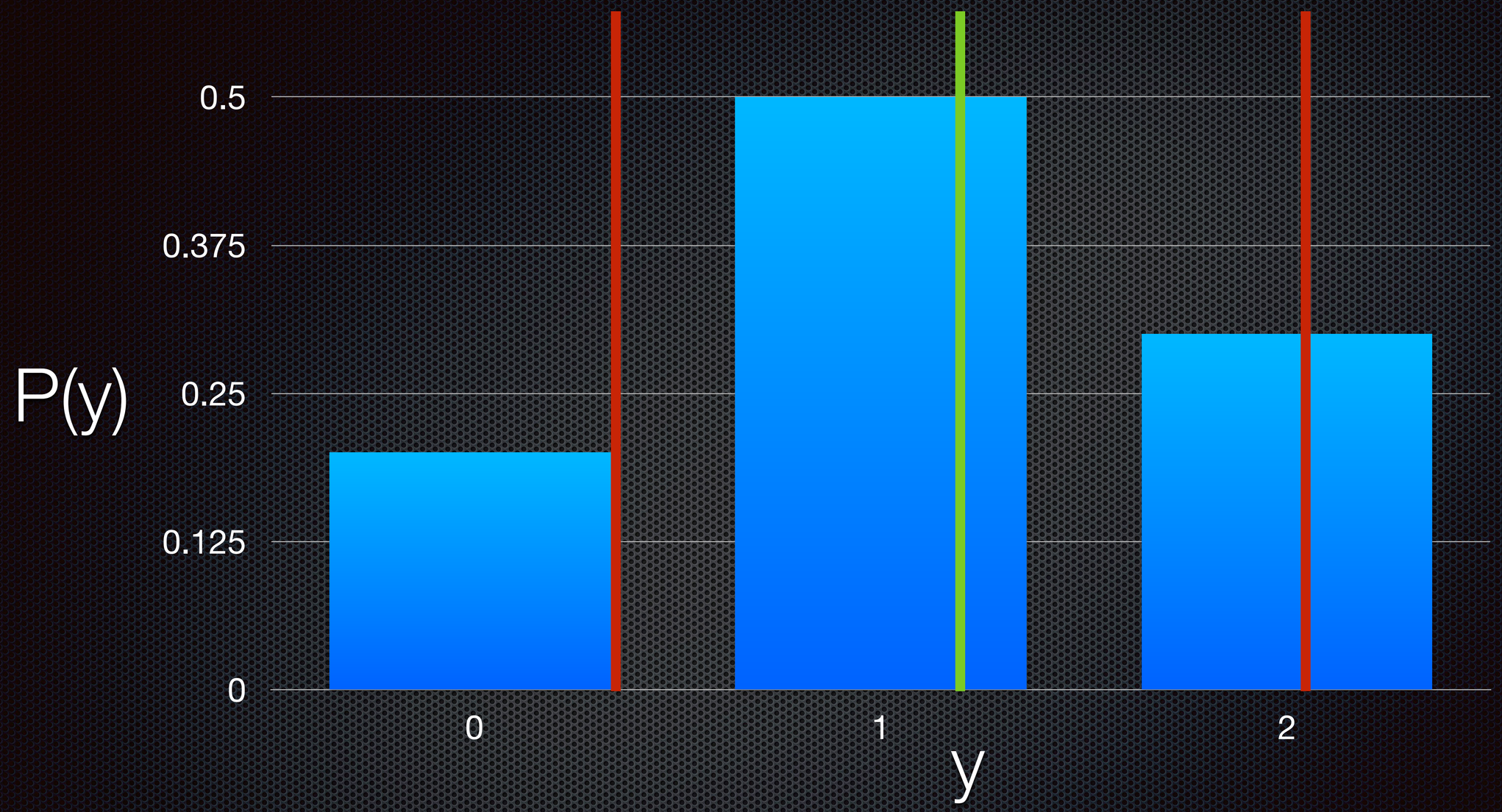


Figure: Histogram representation of the simple distribution including lines at one standard deviation from the mean of the distribution (red lines).

Expectation Computation Example

- Consider the following distribution

y	1	2	3	4
P(y)	0.3	0.2	0.1	0.4

- What is the mean of the distribution
- What is the standard deviation of the distribution?
- Are the mean and standard deviation representative of the distribution form?
- What is the expected value of -logP(y)?

Expectation Example:Answer

- We are given that:

y	1	2	3	4
P(y)	0.3	0.2	0.1	0.4
y^2	1	4	9	16
-log(P(y))	1.204	1.609	2.302	0.916

- Mean:1x0.3+2x0.2+3x0.1+4x0.4=2.6
- Second moment: 1x0.2+4x0.2+9x0.1+16x0.4=8.4
- variance: 8.4-2.6x2.6=1.64
- Standard deviation: $\sqrt{1.64} = 1.2806$
- Expectation -log(P(y)):
  0.3x1.204+0.2x1.609+0.1x2.302+0.4x0.916=1.280

Continuous Variables

- So far discrete values of x or y
- For continuous models we use the probability density function(PDF).
- Discrete case: defined probability distributions over a discrete number of states.
- How do we represent continuous as probability?
- Student heights:
    - Develop a representation which could answer any question we chose to ask about a student's height.
- PDF is a positive function, integral over the region of interest is one.

Manipulating PDFs

Same rules for PDFs as distribution e.g.

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

where p(x,y) =p(x|y)p(y) and for continuous variables $p(x) = \int p(x,y)dy$

Expectations under a PDF

$$< f(x) >_{p(x)} = \int f(x)p(x)dx$$

where the integral is over the region for which our PDF for x is defined.

The Gaussian Density

- Perhaps the most common probability density.

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

- Also available in multivariate form.

- First proposed maybe by de Moire but also used by Laplace.

Gaussian PDF I

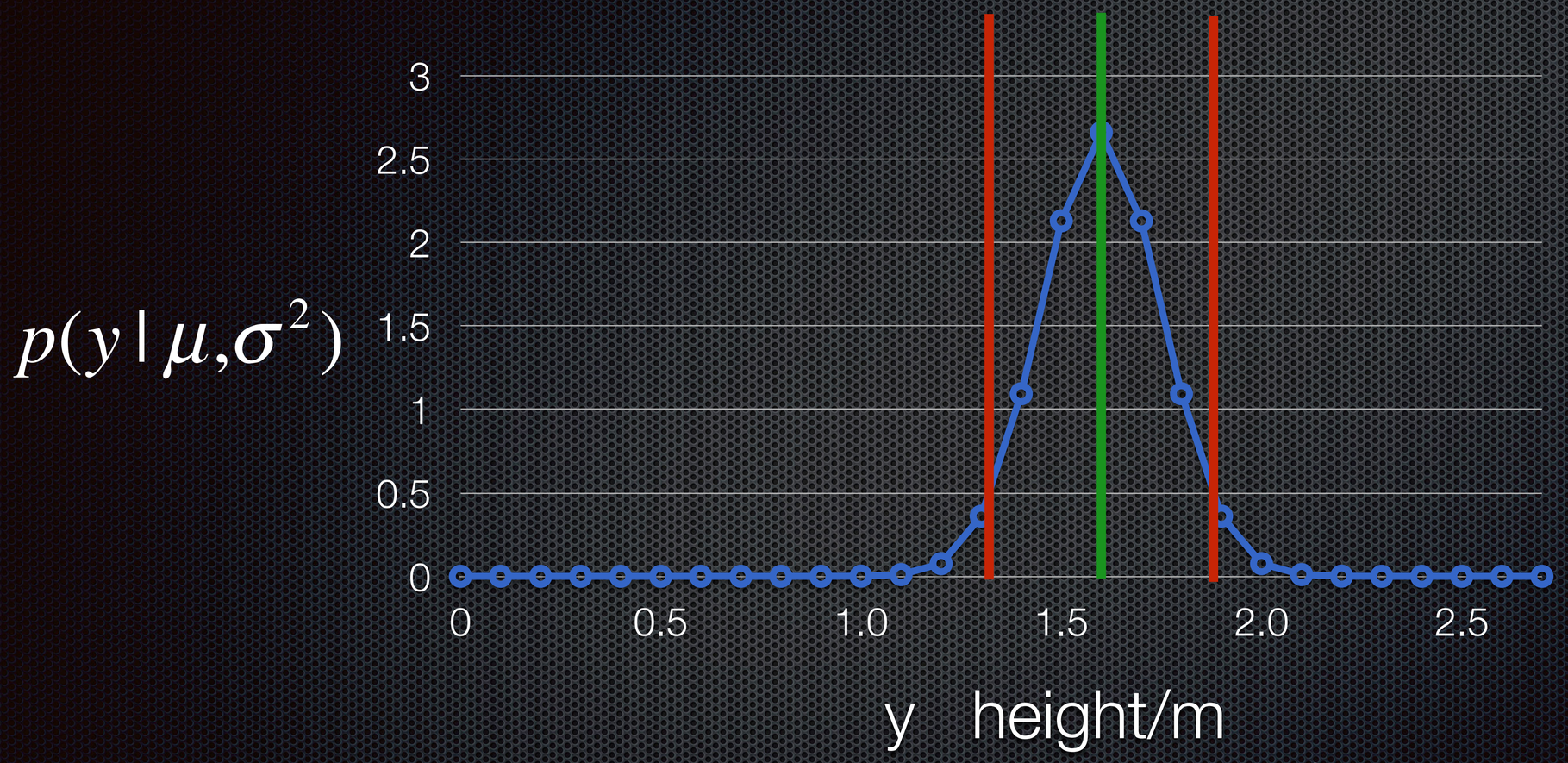


Figure: The Gaussian PDF with μ =1.7 and variance $\sigma^2 = 0.0225$
Mean shows as green line. Two standard deviations are shown as red line.
It could represent the heights of a population of students.

Cumulative Distribution Functions

- PDF doesn't represent probabilities directly
- One very common question is :what is the probability that x<y?
- The cumulative distribution function (CDF) represents the answer

$for -\infty < x < \infty$ the CDF is given by $P(x > y) = \int_{-\infty}^{y} p(x)dx$

$for \ \leq \ 0 \ < \infty$ the CDF is given by $P(x > y) = \int_{0}^{y} p(x)dx$

Gaussian PDF and CDF

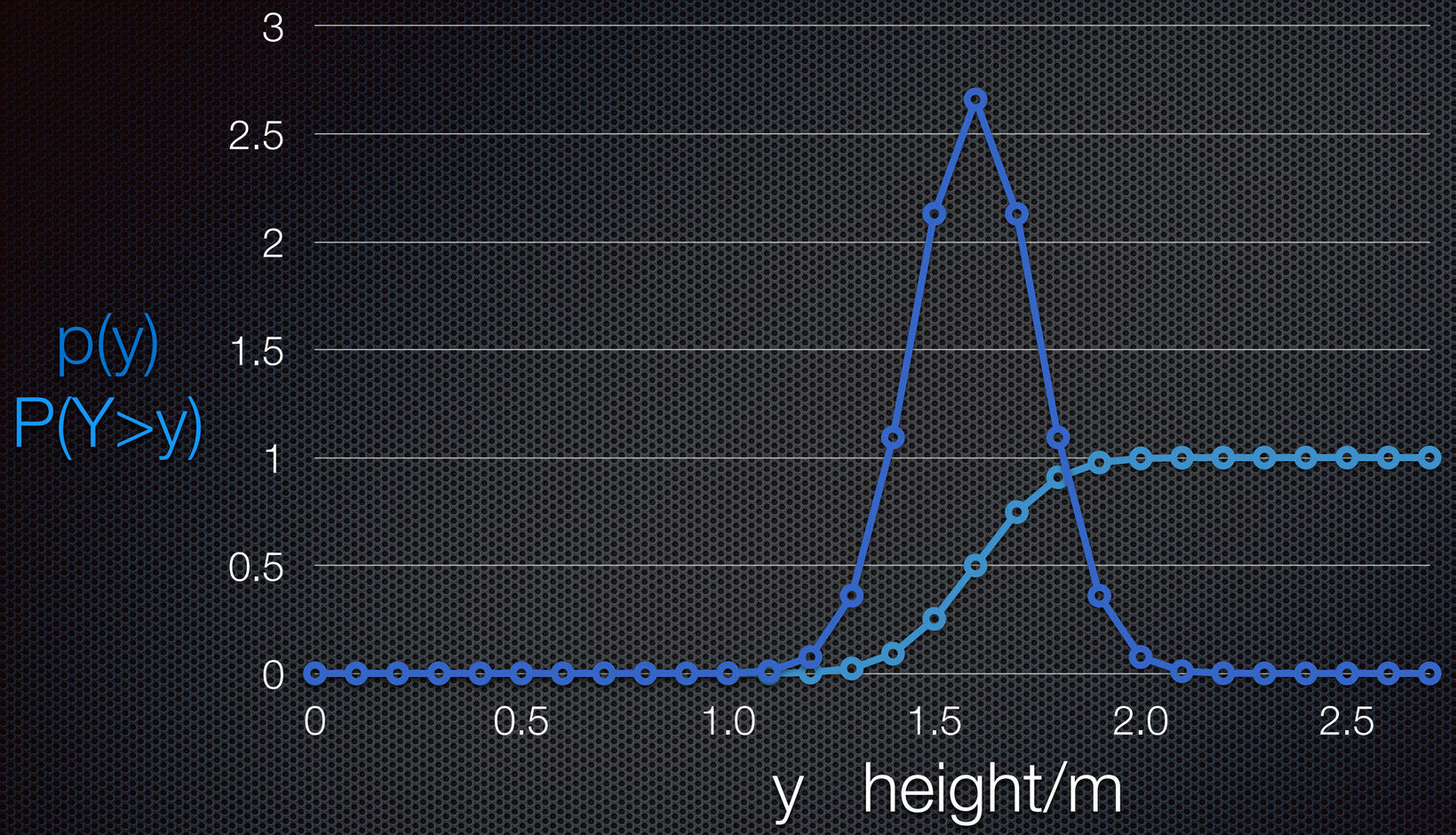


Figure: The Cumulative distribution function(CDF) for the heights of computer science students. The thick curve gives the CDF and the thinner curve the associated PDF.

Sample Based Approximations I

- It is not always possible to compute expectations directly.
- Sample based approximation

$$< f(y) >_{P(y)} \approx \frac{1}{N} \sum_{i=1}^{N} f(y_i)$$

- Special cases of this include the sample mean, often denoted by $\bar{y}$,and computed as

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

Sample Based Approximation Example

- You are given the following values samples of heights of students,

i	1	2	3	4	5	6
y_i	1.76	1.73	1.79	1.81	1.85	1.80

- What is the sample mean?
- What is the sample variance?
- Can you compute sample approximation expected value of -logP(y)?
- Actually these “data” were sampled from a Gaussian with mean 1.7 and standard deviation 0.15. Are you estimates close to the real values? If not why not?

Sample Based Approximation Example:Answer

- We can compute:

i	1	2	3	4	5	6
yi	1.76	1.73	1.79	1.81	1.85	1.80
y_i^2	3.0976	2.9929	3.2041	3.2761	3.4225	3.2400

- Mean: $\frac{1.76+1.73+1.79+1.81+1.85+1.80}{6}=1.79$

- Second moment: $\frac{3.0976+2.9929+3.2041+3.2761+3.4225+3.2400}{6}=3.2055$

- Variance: $3.2055-1.79x1.79=1.43x10^{-3}$

- Standard deviation 0.0379

No, you can't compute it. You don't have access to P(y) directly.

Sample Mean vs True Mean

- This is an approximation to the true distribution mean

$$< y > \approx \bar{y}$$

- The same approximation can used for continuous PDFs, so we have

$$< f(y) >_{p(y)} = \int f(y) p(y) dy \approx \frac{1}{N} \sum_{i=1}^{N} f(y_i)$$

where y_i are independently obtained samples from the density p(y).

Approximation gets better for increasing N and worse if the samples from P(y) are not independent.

Adapting Maximum Likelihood approach

- In the standard Gaussian, parametrized by mean and variance.
- Make the mean a linear function of an input.
- This leads to a regression model.

$$t_i = y(x_i) + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

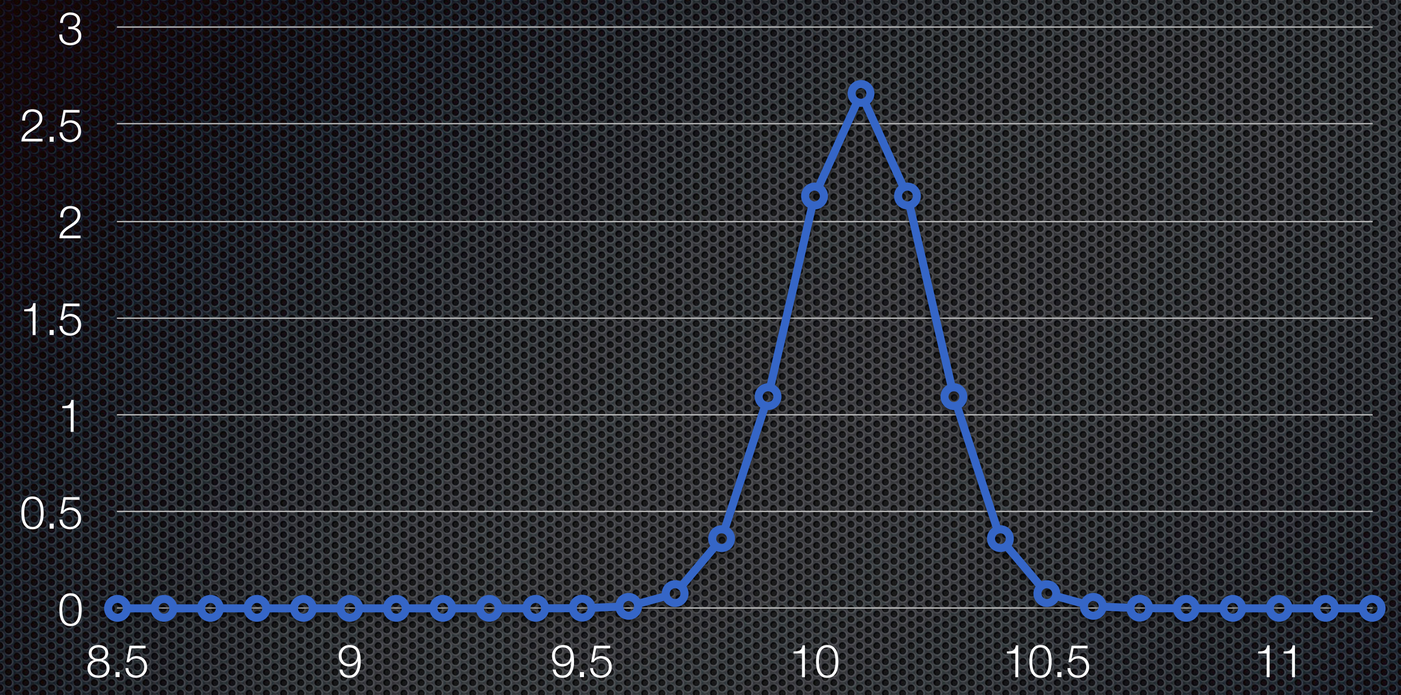
- t_i is time and x_i is year. We take men's 100m records of Summer Olympics Games from 1896 to 2008 as an example.

Adapting Maximum Likelihood approach

$$t_n = f(x_n; w) + \varepsilon_n, \quad \varepsilon_n \sim N(0, \sigma^2).$$

$$p(t_n \mid x_n, w, \sigma^2) = N(w^T x_n, \sigma^2)$$

Adapting Maximum Likelihood approach

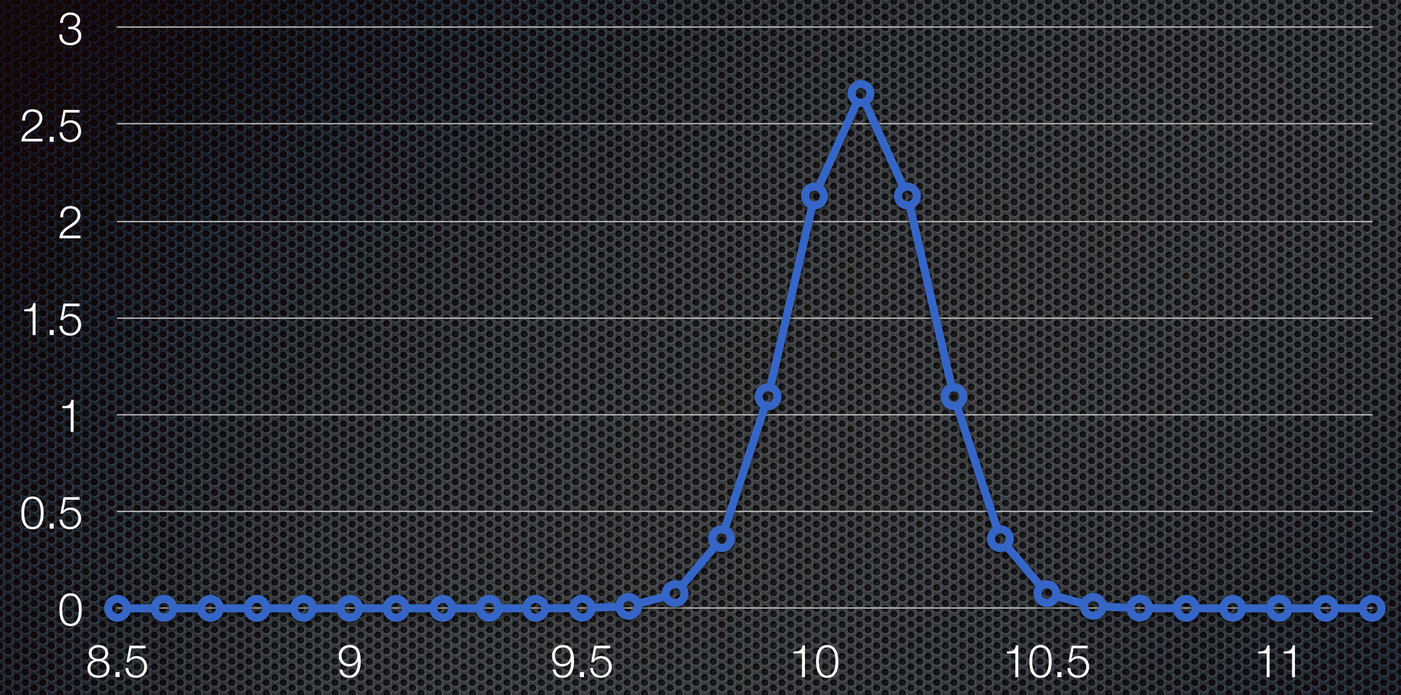


$$p(t_n \mid x_n = [1,1980]^T, w = [36.416, -0.0133]^T, \sigma^2 = 0.05)$$

$$\mu = 36.416 - 0.0133 \times 1980 = 10.02$$

10.25 actual time in 1980

Adapting Maximum Likelihood approach



Change w and σ^2 to try and move the density so as to make it as high as possible at t=10.25

Dataset likelihood

Joint conditional density

$$p(t_1,\ldots,t_N \mid x_1,\ldots,x_N,w,\sigma^2)$$

$$L = p(t \mid X,w,\sigma^2) = \prod_{n=1}^{N} p(t_n \mid x_n,w,\sigma^2) = \prod_{n=1}^{N} N(w^T x_n,\sigma^2)$$

Maximum likelihood

Natural logarithm

$$\log L = \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(t_n - f(x_n;w))^2\right\}\right)$$

$$= \sum_{n=1}^{N} \left(-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}(t_n - f(x_n;w))^2\right)$$

$$= -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - f(x_n;w))^2$$

Maximum likelihood

Taking derivatives, equating them to zero and solving for turning points.

$$\log L = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - w^T x_n)^2$$

$$\frac{\partial \log L}{\partial w} = \frac{1}{\sigma^2}\sum_{n=1}^{N} x_n(t_n - x_n{}^T w) = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n t_n - x_n x_n{}^T w)$$

$$= \frac{1}{\sigma^2}(X^T t - X^T X w) = 0$$

Maximum likelihood

Solution:

$$\frac{1}{\sigma^2}(X^T t - X^T X w) = 0$$

$$X^T t - X^T X w = 0$$

$$X^T t = X^T X w$$

$$\hat{w} = (X^T X)^{-1} X^T t$$

Maximum likelihood

Solution for the noise variance:

$$\frac{\partial L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{n=1}^{N}(t_n - x^T\widehat{w})^2 = 0$$

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - x^T\widehat{w})^2$$

$$= \frac{1}{N}(t - X\widehat{w})^T(t - X\widehat{w})$$

$$= \frac{1}{N}(t^T t - 2t^T X\widehat{w} + wX^T X\widehat{w})$$

Maximum likelihood

Solution for the noise variance:

$$\widehat{\sigma^2} = \frac{1}{N}(t^T t - 2t^T X(X^T X)^{-1} X^T t + (X^T X)^{-1} X^T t X^T X(X^T X)^{-1} X^T t)$$

$$= \frac{1}{N}(t^T t - 2t^T X(X^T X)^{-1} X^T t + t^T X(X^T X)^{-1} X^T t)$$

$$= \frac{1}{N}(t^T t - t^T X(X^T X)^{-1} X^T t)$$

$$\widehat{\sigma^2} = \frac{1}{N}(t^T t - t^T X\widehat{w})$$