Mathematics in ML

1856 १८५६

उत्तिष्ठत जाग्रत प्राप्य वरान् निबोधत

INDIAN INSTITUTE OF ENGINEERING SCIENCE AND TECHNOLOGY, SHIBPUR

भारतीय अभियांत्रिकी विज्ञान एवं प्रौद्योगिकी संस्थान, शिवपुर

Jaya Sil

Department of Computer Science and Technology

Topics

- Calculus is used for **optimizing** the error or loss.
- Linear Algebra is used to transform equation into matrix
- Probability is measure of **uncertainty** and **likelihood e**stimation.
- Logistic regression, Multivariate regression for **classification**

Uncertainty

- Machine learning deals with uncertain quantity and stochastic (non-deterministic) quantity because all activities require some ability to reason in the presence of uncertainty.
- Inherent stochasticity in the system being modeled.
- Incomplete observability, when we cannot observe all of the variables that model the behavior of the system.
- Incomplete modeling when we use a model that must discard some of the information we have observed, the discarded information results in uncertainty in the model's predictions.

Why probability ?

- Probability theory was originally developed to analyze the frequencies of events where events are often repeatable, like tossing a coin, drawing a card.

- This kind of reasoning does not seem immediately applicable to propositions that are not repeatable.

- Probability to represent *degree of belief* with 1 indicates absolute certainty and 0 indicating absolute certainty of earlier observation.

- Probability, related directly to the rates at which events occur, is known as *frequentist probability, while related to qualitative levels of* certainty, is known as *Bayesian probability.*

Why Probability?

- Probability can be seen as the extension of logic to deal with uncertainty.

- Logic provides a set of formal rules for determining what propositions are implied to be true or false given the assumption that some other set of propositions is true or false.

- Probability theory provides a set of formal rules for determining the likelihood of a proposition being true given the likelihood of other propositions.

Random Variable

- A random variable is a variable that can take on different values randomly.

- The set of all possible outcomes and their probabilities is known as probability distribution, describes how the total probability (=1) is distributed over all possible outcomes.

- A random variable describing the states that are possible and probability distribution specifies how likely each of these states are.

- Random variables may be discrete or continuous.

Discrete Variables and Probability Mass Functions

- Probability distribution over discrete variables is described using *probability mass function (*PMF).
- The PMF maps from a state of a random variable to the probability of that random variable taking on that state.
- PMF is written using the random variable: P (X= *x*).
- PMF can act on many variables at the same time.
- Probability distribution over many variables is known as joint probability distribution where *P*(X = *x,* Y = *y*) denotes the probability that X = *x* and Y= *y* occur simultaneously.

Continuous Variables and Probability Density Functions

- Probability distributions of continuous random variables are described using a *probability density function* (*PDF*).

- A probability density function p(*x*) does not give the probability of a specific state directly, the probability of inside an infinitesimal region with volume δ*x* given by p(*x*)δ*x*.

$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x.$$

$$\begin{aligned} p(x) &\geqslant 0 \\ \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x &= 1. \end{aligned}$$

In the univariate example, the probability that *x* lies in the interval [a, b].

Marginal Probability

- The probability distribution over the subset is known as the *marginal probability distribution.*

From joint PMF, we find individual PMF called marginal PMF.

suppose x and y are discrete random variables, and P (x, y) is known.

We find P (x) with the *sum rule:* $\forall x \in \mathrm{x}, \mathrm{P}(\mathrm{x} = x) = \sum_{\mathrm{y}} \mathrm{P}(\mathrm{x} = x, \mathrm{y} = y)$

For continuous random variable we apply integration over *pdf*

$$\forall x \in \mathrm{x}, \mathrm{P}(\mathrm{x} = x) = \int \mathrm{P}(\mathrm{x} = x, \mathrm{y} = y) \mathrm{d}y$$

Conditional Probability

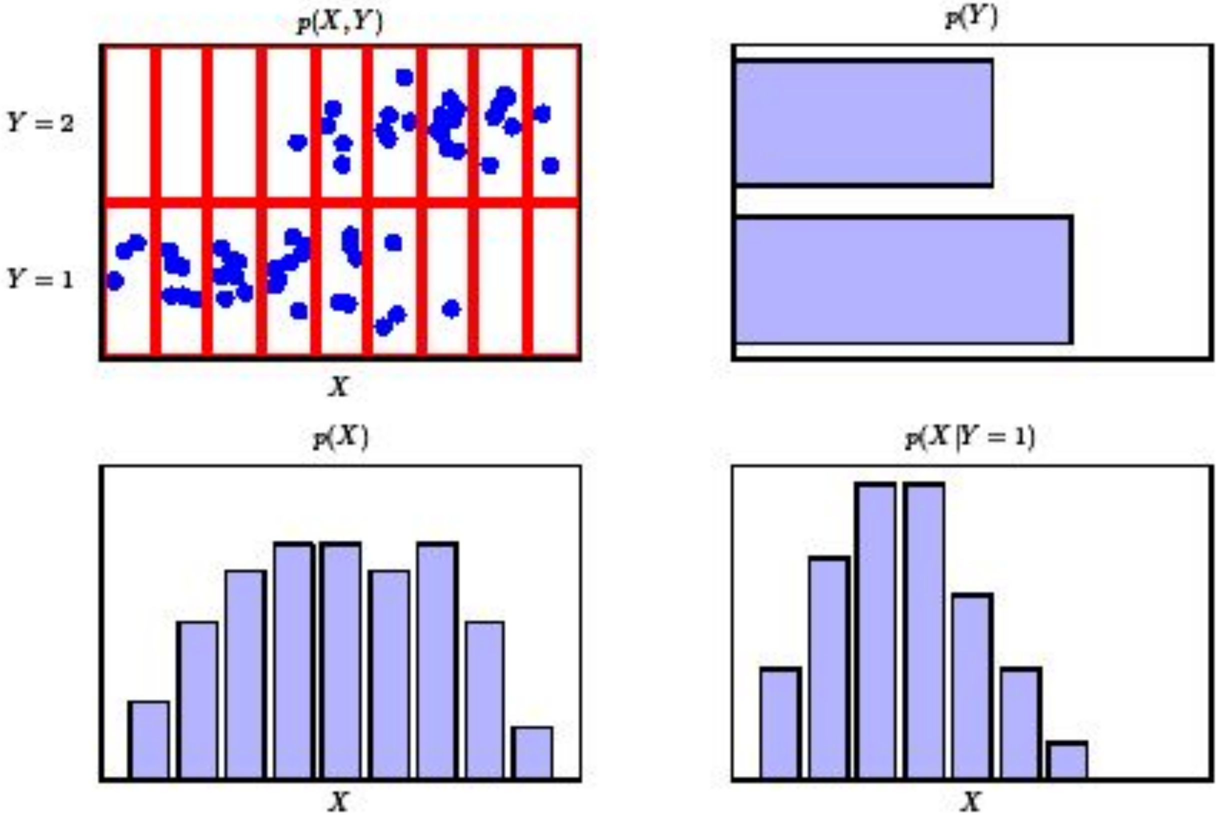
- Conditional probability of event y = *y* given that event x = *x* already occurred, defined by $$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$$

- Conditional probability is useful when given observed events, our belief or prediction of related events can be updated.

- $p(x \cap y) = p(x|y)\, p(y)$ also $p(x \cap y) = p(y|x)\, p(x)$
- $p(x|y) = p(y|x)\, p(x) / p(y)$

- We cannot compute the conditional probability conditioned on an event that never happens.

Conditional and Marginal Probability Distribution



p(X, Y) = p(X) p(Y); X and Y are called independent

The Chain Rule of Conditional Probabilities

- Joint probability distribution over many random variables may be decomposed into conditional distributions over only one variable

$$P(\mathrm{x}^{(1)}, \dots\dots \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)})\prod_{i=2}^{n} P(\mathrm{x}^{(i)}|\mathrm{x}^{(1)} \dots\dots \mathrm{x}^{(i-1)}$$

This observation is known as the chain rule or product rule of probability

- P (a, b, c) = P(a| b, c) P(b,c)

- P(b, c) = P(b|c) P(c)

- P(a,b,c) = P(a| b, c) P(b|c) P(c)

- Apply conditional probability to obtain **Bayes' Rule**!

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

Conditioned Bayes' Rule: given events A;B; C

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

Independence and Conditional Independence

- Two events x and y are said to be independent if $\forall x \in$ x, $y \in$ y, p(x=x, y=y) = p(x=x) p(y=y)

  The probability that one event occurs in no way affects the probability of the other event occurring.

- Two random variables x and y are conditionally independent given a random variable z if the conditional probability distribution over x and y factorizes for every value of z:

$\forall x \in$ x, $y \in$ y, $z \in$ z p(x=x, y=y|z=z) = p(x=x|z=z) p(y=y|z=z)

Expectation

- The *expectation or expected value* of function $f\,(x)$ with respect to a probability distribution P (x) is the average or mean value that f takes on when x is drawn from P .

- For discrete variables: $\mathrm{E}_{\mathrm{x\sim P}}\,[f\,(x)] = \sum_{x} \mathrm{P}(x)\,f\,(x)$

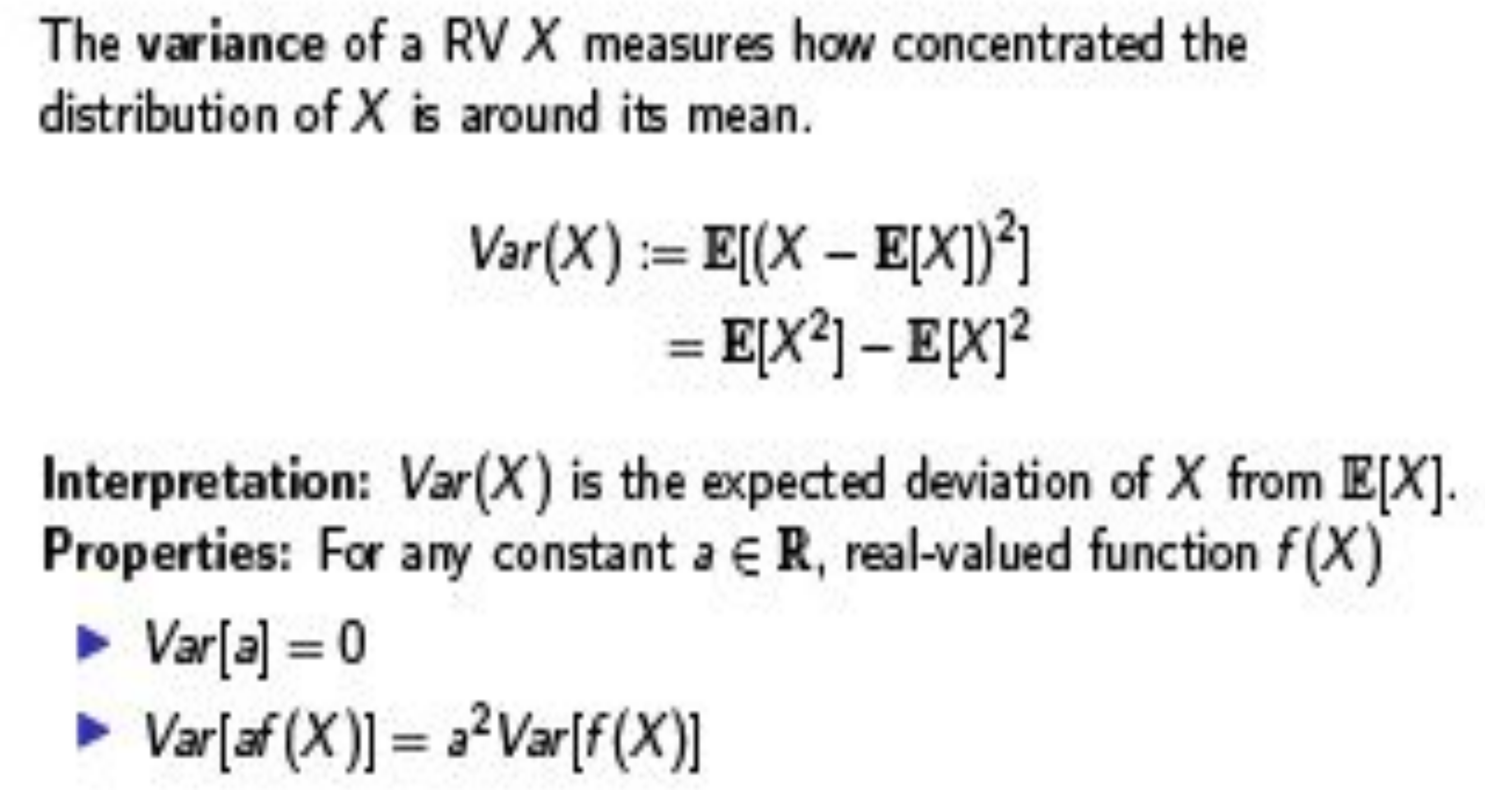
- For continuous variables: $\mathrm{E}_{\mathrm{x\sim P}}\,[f\,(x)] = \int \mathrm{P}(x)\,f\,(x)\;\mathrm{d}x$

Expectation of Random Variable

- The random variable X takes value as -2, -1, 1, 3 with probabilities ¼, 1/8, ¼ and 3/8, respectively. What is the expectation of random variable $Y = X^2$?

- $E(Y) = E(X^2) = 4×1/4 +1 × 1/8 + 1 × 1/4 + 9 × 3/8 = 19/4$

Variance



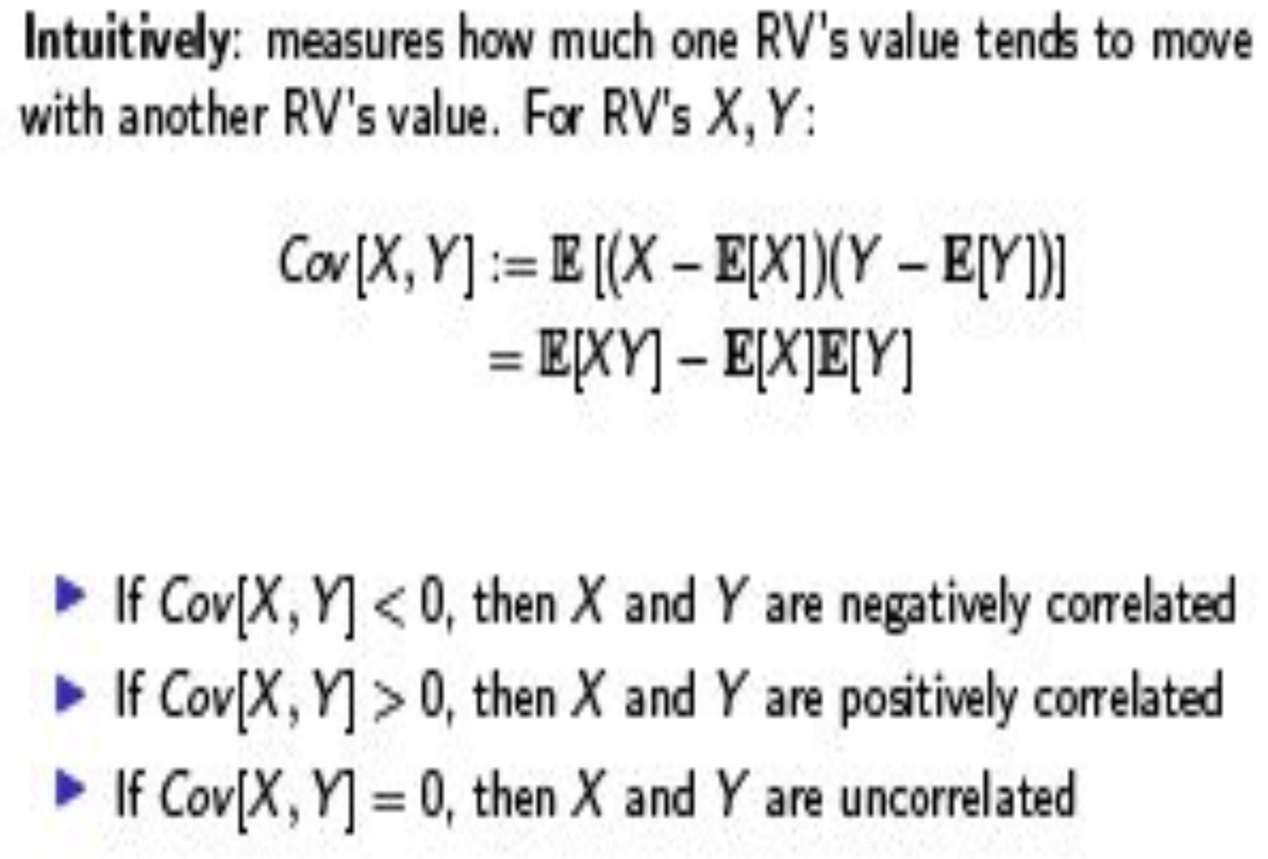
The **variance** of a RV X measures how concentrated the distribution of X is around its mean.

$$Var(X) := \mathbf{E}[(X - \mathbf{E}[X])^2]$$
$$= \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

Interpretation: $Var(X)$ is the expected deviation of X from $\mathbb{E}[X]$.
Properties: For any constant $a \in \mathbf{R}$, real-valued function $f(X)$

- $Var[a] = 0$
- $Var[af(X)] = a^2 Var[f(X)]$

Covariance

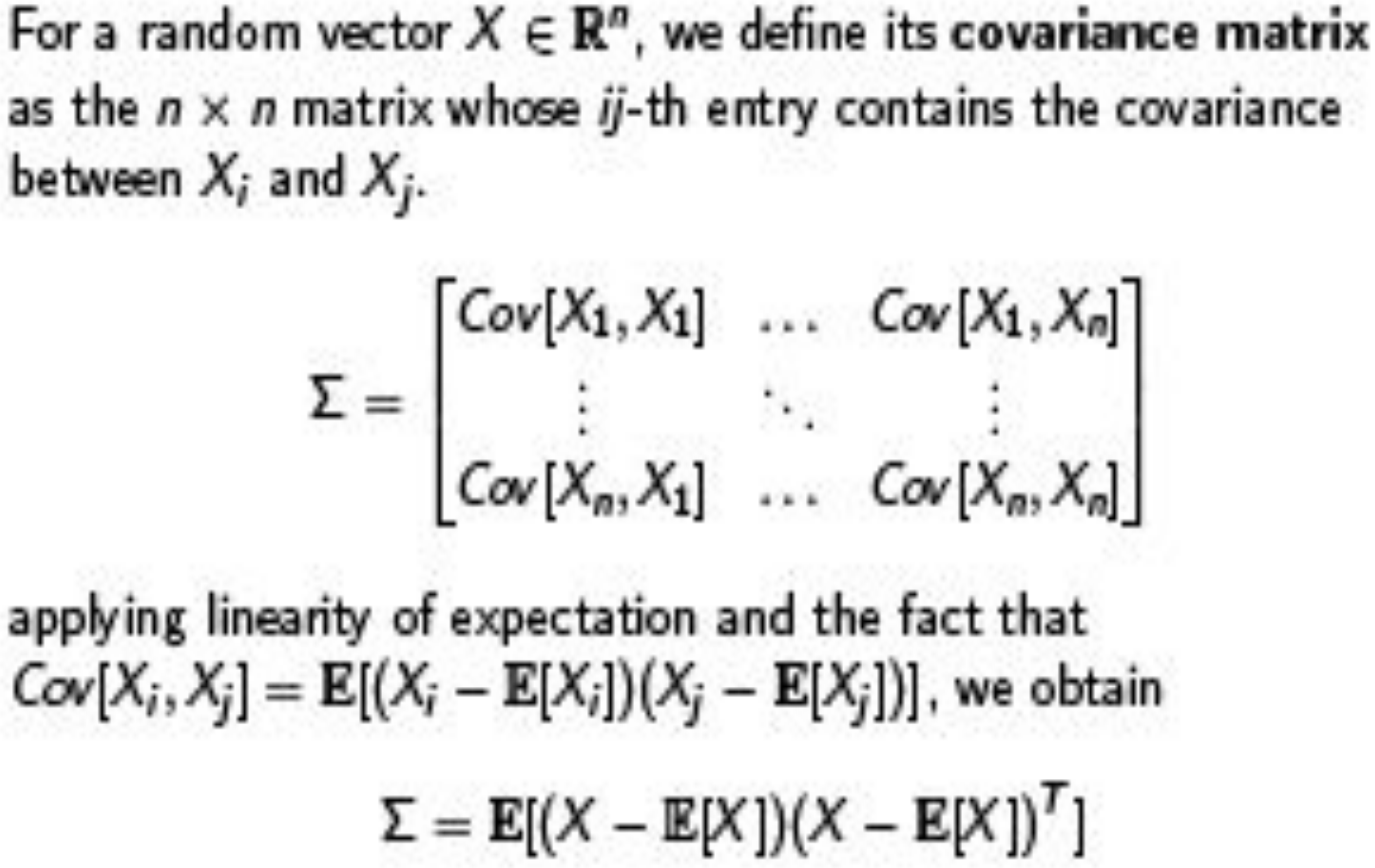


Intuitively: measures how much one RV's value tends to move with another RV's value. For RV's X, Y:

$$\begin{aligned} Cov[X,Y] &:= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

- If $Cov[X,Y] < 0$, then X and Y are negatively correlated
- If $Cov[X,Y] > 0$, then X and Y are positively correlated
- If $Cov[X,Y] = 0$, then X and Y are uncorrelated

Covariance Matrix



For a random vector $X \in \mathbf{R}^n$, we define its **covariance matrix** as the $n \times n$ matrix whose ij-th entry contains the covariance between X_i and X_j.

$$\Sigma = \begin{bmatrix} Cov[X_1, X_1] & \dots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \dots & Cov[X_n, X_n] \end{bmatrix}$$

applying linearity of expectation and the fact that $Cov[X_i, X_j] = \mathbf{E}[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])]$, we obtain

$$\Sigma = \mathbf{E}[(X - \mathbf{E}[X])(X - \mathbf{E}[X])^T]$$

Likelihood Function

- $p(D|\boldsymbol{w})$ is evaluated for the observed data set D as a function of the parameter vector **w**, it is called the *likelihood* function**.**

- It expresses how probable the observed data set is for different settings of the parameter vector **w**.

- *likelihood* is not a probability distribution over **w,** and its integral with respect to **w** does not (necessarily) equal one.

- posterior $\propto$ *likelihood* × *prior*

- All quantities are viewed as functions of **w.**

Maximum Likelihood

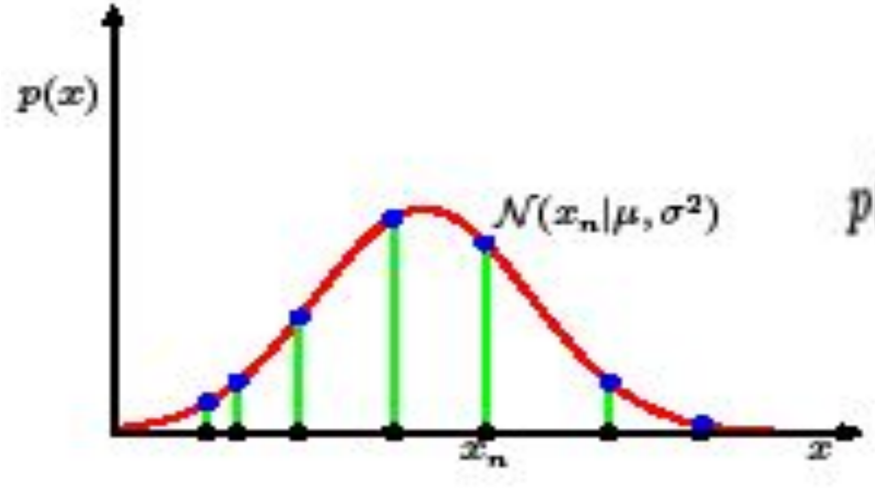
- Data set consisting of Continuous independent variables and a continuous dependent variable.

- A widely used estimator is *maximum likelihood*, in which $\boldsymbol{w}$ is set to the value that maximizes the likelihood function $p(D|\boldsymbol{w})$.

- In the machine learning literature, the negative log of the *likelihood* function is called an *error function.*

- The negative logarithm is a monotonically decreasing function, maximizing the *likelihood* is equivalent to minimizing the *error function = - log(y)*

Maximizing the likelihood



$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right).$$

- Black points denote a dataset and the likelihood function corresponds to the product of the blue values.

- Maximizing the likelihood involves adjusting the mean and variance of the Gaussian so as to maximize this product.

Maximizing w.r.t. parameters

Maximizing with respect to *μ,* we obtain the maximum likelihood, i.e., the mean of the observed values $\{x_n\}$.

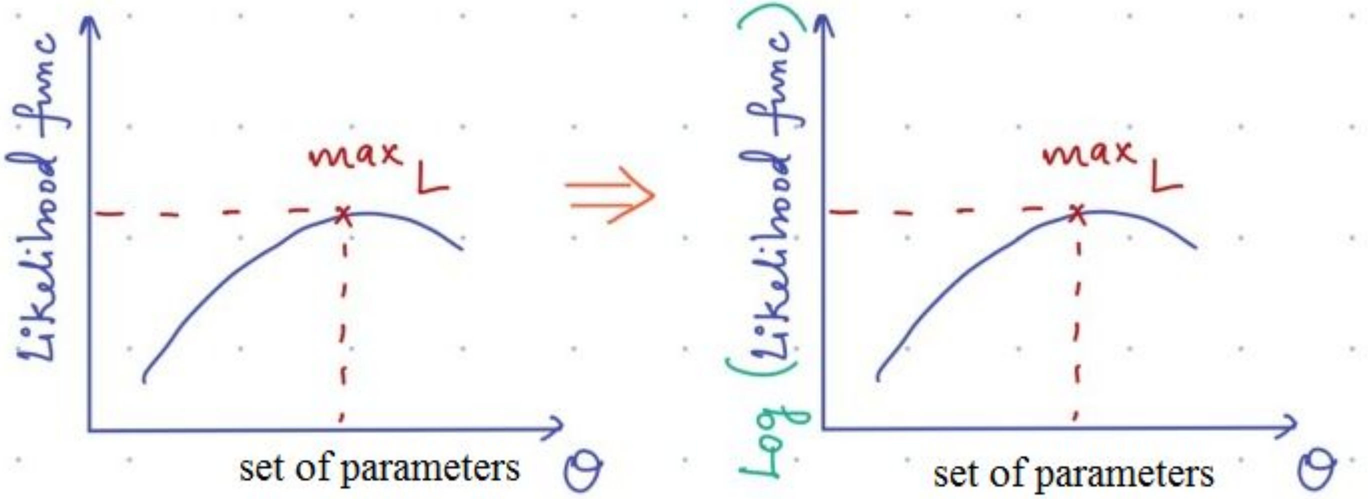
$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Similarly, maximizing with respect to σ^2*,* we obtain the maximum likelihood solution for the variance in the form

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$$

Log Likelihood function

$$\log(L) = \log f(x_1|\theta) + \log f(x_2|\theta) + \ldots + \log f(x_N|\theta)$$



Applying log to the likelihood function simplifies the expression into a sum of the log of probabilities and does not change the graph with respect to θ

The Gaussian distribution

- a single real valued variable *x*, the Gaussian distribution is:

$$
\begin{aligned}
N(x|\mu,\, \sigma^2) &= p(x;\; \mu,\sigma^2) \\
&= \frac{1}{(2\Pi)^{1/2}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)
\end{aligned}
$$

For The reciprocal of the variance, written as $\beta = 1/\sigma^2$, is called the precision.

$$\int_{-\infty\,\ldots\,+\infty} N\,(x|\,\mu,\, \sigma^2)\;\mathrm{d}x = 1$$

Expectations of functions under the Gaussian distribution

Expectations of functions of *x* under the Gaussian distribution, i.e. the average value of *x* is given by

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \, \mathrm{d}x = \mu.$$

The second order moment: $$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2.$$

The variance of *x* is $$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

The Multivariate Gaussian Distribution

- A vector-valued random variable $X = [X_1 \cdot \cdot \cdot X_n]^T$ is said to have a multivariate normal (or Gaussian) distribution with mean $\mu \in R^n$ and covariance matrix $\sum$, if its probability density function is given by

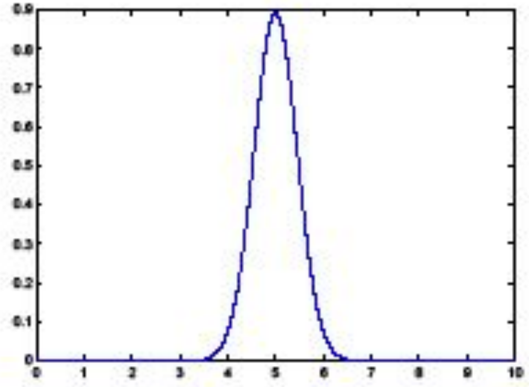
$$p(x;\ \mu, \Sigma) = \frac{1}{(2\Pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^{T}\Sigma^{-1}(x-\mu)\right)$$

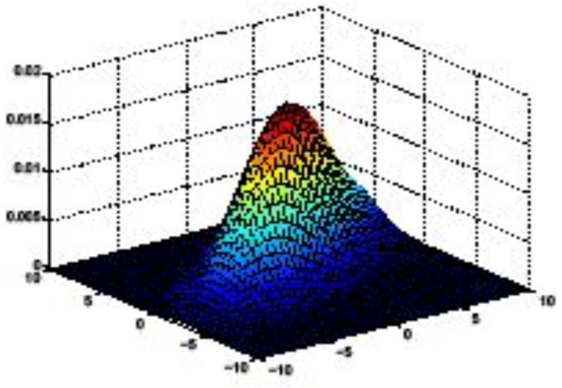
pdf is also given by $N(x \mid \mu, \sum)$

$\sum$ must be symmetric positive definite matrix

- A univariate Gaussian density for a single variable X

- A multivariate Gaussian density over two variables X_1 and X_2.





The covariance matrix

$$p(x|\mu,\Sigma) = \frac{1}{(2\Pi)^{1/2}\sigma_1}\exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2\right)\cdot\frac{1}{(2\Pi)^{1/2}\sigma_2}\exp\left(-\frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2\right)$$

- product of two independent Gaussian densities, one with mean μ_1 and variance σ_1 and the other with mean μ_2 and variance σ_2

- More generally, one can show that an n-dimensional Gaussian with mean $\mu \in R^n$ and diagonal covariance matrix $\sum = \mathrm{diag}(\sigma_1^2, \sigma_2^2 \ldots \sigma_n^2)$ is the same as a collection of *n* independent Gaussian random variables with mean μ_i and variance σ_i^2, respectively.

Interpretation

- Covariance matrix provides us two useful information.
- The diagonal elements tell us how much variability we might expect in the individual parameters-how well they are defined by the data.
- If parameters vary quite a lot then they are not defined very well by the data.
- Off-diagonal elements tell us how the parameters co-vary-if the values are high and positive, increasing one will require increase of another to maintain a good model.

independent and identically distributed

- A data set of observations $\mathbf{x} = (x_1, \ldots, x_N)^T$, representing N observations of variable x.
- The observations are drawn independently from a Gaussian distribution whose mean μ and variance σ^2 are unknown.
- Determine the parameters mean and variance.
- Data points that are drawn independently from the same distribution are said to be *independent and identically distributed.*

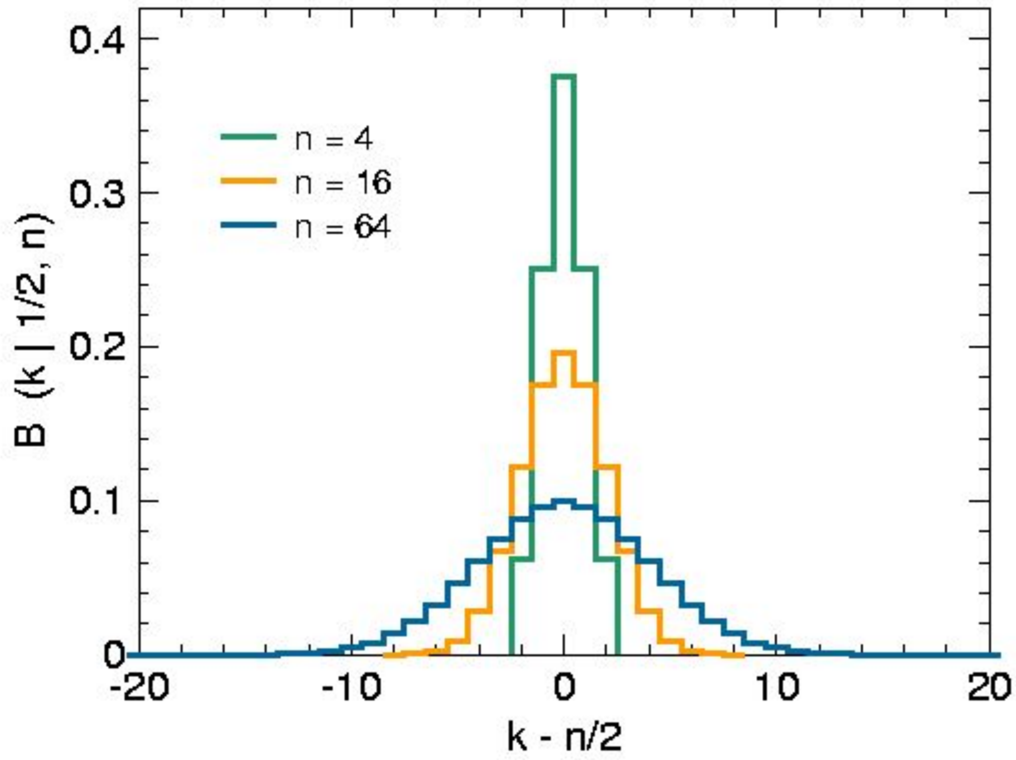
Bernoulli Trial

- A **Bernoulli Trial** is an experiment whose outcome is random (and has one of two outcomes (e.g. heads or tails).
- PMF of x; p(x=1) = p and p(x=0) =1- p
- *E*[x] = *p, var*(x) = *p* (1-*p*)
- Think of it is a Boolean random variable, ***x*.**
- A set of random {***x*1, *x*2 ,….*xn***} variables is *independent and identically distributed* (IID) if all variables in the set are mutually independent and all are governed by the same probability distribution ***D*.**

Example: coin flip

- Assume an unbiased coin X that takes two values {0,1}.
- If all coin flips use the same coin, we assume that they are IID Bernoulli trials
- This is modelled by the Binomial Distribution when we perform multiple Bernoulli trials.
- $E[X] = np$, $var(X) = np\,(1-p)$
- As ***n*** goes to infinity, the Normal distribution approximates the Binomial distribution

- Classification is like a coin flip, you’re either right or wrong.
- If classification is independent, then the number of correct classifications is governed by a Binomial distribution.
- Binomial distribution is approximated by the Normal distribution.
- The Normal distribution lets us estimate how close the TRUE error is to the SAMPLE error.



the number of correct classifications ***K***

Probabilistic Interpretation

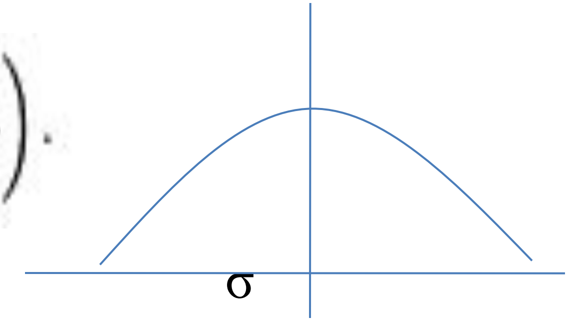
- The relation between the target variables and the inputs:
  $y^{(i)} = W^T x^{(i)} + \varepsilon^{(i)}$ (*Error: Random noise*)

- $\varepsilon^{(i)}$ is an error term that captures selection of features

$\varepsilon^{(i)} \sim N(0, \sigma^2)$; $\varepsilon^{(i)}$ is *iid* and represented using a Normal distribution with mean zero and variance σ^2.

- Density of error:

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right).$$



$$p(y^{(i)}|x^{(i)}; W) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - W^T x^{(i)})^2}{2\sigma^2}\right).$$

ERROR Distribution

$p(y^{(i)}|x^{(i)};w)$ indicates that this is the distribution of $y^{(i)}$ given $x^{(i)}$ and parameterized by w.

- Distribution of $y^{(i)}$ is written as $N\left(w^T x^{(i)}, \sigma^2\right)$ for a fixed value of w

- When we wish to explicitly view this as a function of w, we will instead call it the likelihood function: $L(w) = L(w;X,y) = p(y|x;w)$

- The principal of maximum likelihood states that we choose w so as to make data as high probability as possible.

- It means we choose w so that L(w) is maximum

Maximum Likelihood

- Based on the *iid* assumption of $\varepsilon^{(i)}$, true for $y^{(i)}$ and $x^{(i)}$
- $$L(w) = \prod_{i=1}^{n} p(y^{(i)}|x^{(i)};w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\Pi\sigma^2}} exp\left(-\frac{(y^{(i)} - w^T x^{(i)})^2}{2\sigma^2}\right)$$
- Instead of maximizing L(w), we can maximize strictly increasing function of L(w), i.e. the log likelihood ℓ(w)
- ℓ(w) = log L(w) = $$log \prod_{i=1}^{n} \frac{1}{\sqrt{2\Pi\sigma^2}} exp\left(-\frac{(y^{(i)} - w^T x^{(i)})^2}{2\sigma^2}\right)$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\Pi\sigma^2}} exp\left(-\frac{(y^{(i)} - w^T x^{(i)})^2}{2\sigma^2}\right) = nlog \frac{1}{\sqrt{2\Pi}} \frac{1}{\sigma} - \frac{1}{\sigma^2} . \frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - w^T x^{(i)})^2$$

Maximizing ℓ(w) same as minimizing $\frac{1}{2}\sum_{i=1}^{n} (y^{(i)} - w^T x^{(i)})^2$ i.e. Least-Square Cost function

- Least-squares regression corresponds to finding the maximum likelihood estimate of w.

- Final choice of w does not depend on what is σ^2, we arrive at the same result even if σ^2 are unknown.

- So in Linear regression problem the least-squares cost function L, be a reasonable choice.

NORMS

- We measure the length of a vector using a function called *Norm*.
- *Norm* functions mapping vectors to nonnegative values.
- Usually the Norm of a vector **x** measures the distance from the origin to the point **x**.
- A Norm is a function *f* that satisfies the following properties:
- $f(\mathbf{x}) = 0$; $\mathbf{x} = \mathbf{0}$
- $f(\mathbf{x+y}) \le f(\mathbf{x}) + f(\mathbf{y})$
- For all α ε R, $f(\alpha\mathbf{x}) = |\alpha|\, f(\mathbf{x})$

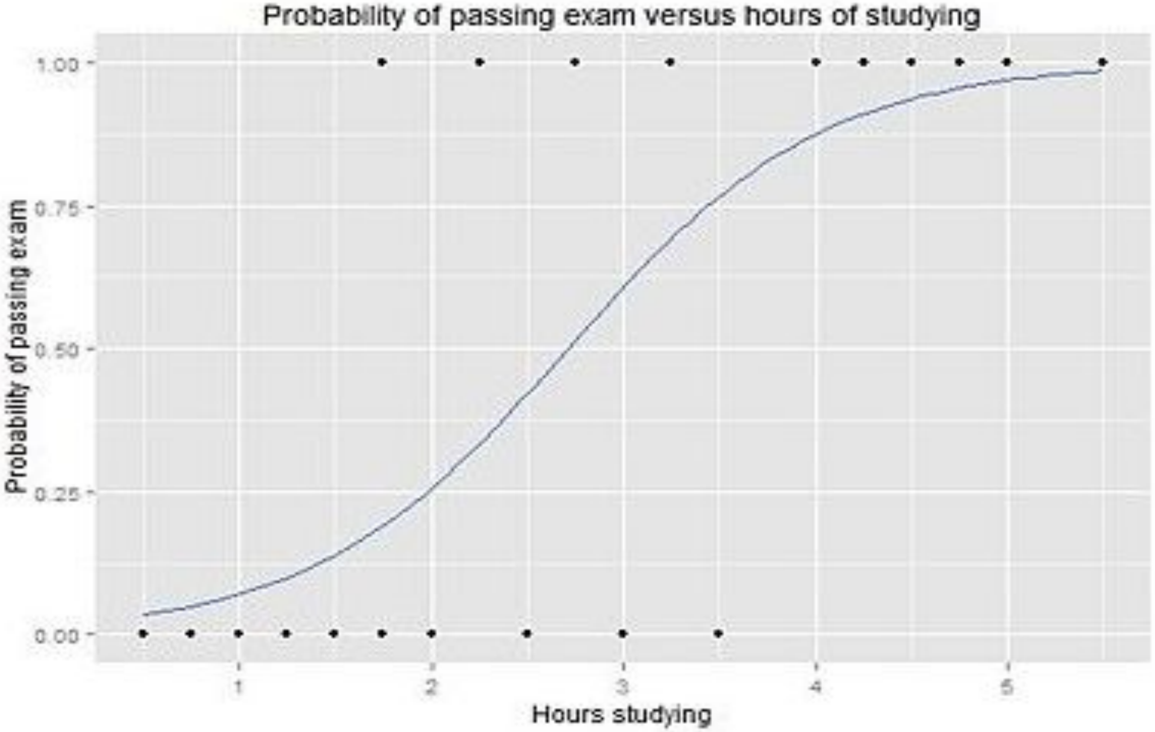
Norm

- The L^p norm is $||\mathbf{x}||_p = (\Sigma|\mathbf{x}_i|^p)^{1/p}$; p > = 1
- L^2 Norm (p =2) known as Euclidean, denoted as $||x|| = (\Sigma x_i^2)^{1/2}$
- L^2 Norm, written as $||x||_2 = x^T x$
- L^1 Norm $||\mathbf{x}||_1 = \Sigma|x_i|$
- Max Norm $L^\infty = \Sigma||\mathbf{x}||_\infty = \max_i |x_i|$
- Size of a matrix using Frobenious Norm $||\mathbf{A}||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2}$
- Dot product of two vectors $x^T y = ||x||_2 \, ||y||_2 \cos \theta$

Logistic Regression

- Logistic regression is a regression model used for classification and predicting a discrete variable.
- The coefficients in logistic regression are estimated using a process called maximum-likelihood estimation.
- In logistic regression we use a hypothesis class to predict the probability that a given example belongs to the “1” class versus the probability that it belongs to the “0” class.
- Specifically, we learn a function σ() of the form:
- $h_w(x) = P(y = 1|x) = 1 / 1+\exp(-\mathbf{w}^\top \mathbf{x}) \equiv \sigma(\mathbf{w}^\top \mathbf{x})$
- $P(y = 0|x) = 1- P(y=1|x) = 1-h_w(x)$

- The graph shows the probability of passing the exam versus the number of hours studying, with the logistic regression curve fitted to the data.



- Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables.

Logistic Function

- The function $\sigma(z) \equiv 1 / 1 + \exp(-z)$ is often called the "sigmoid" or "logistic" function where $z = \mathbf{w}^\top \mathbf{x}$

- It is an S-shaped function that keep the value of $\mathbf{w}^\top \mathbf{x}$ into the range [0,1] so that we may interpret hypothesis $h_w(x)$ as a probability.

- Our goal is to search for a value of *w* so that the probability $P(y = 1|x) = h_w(x)$ is large when *x* belongs to the "1" class and small when *x* belongs to the "0" class (so that $P(y=0|x)$ is large).

Cost or Loss Function

- For a set of training examples with binary labels $\{(x^{(i)}, y^{(i)}): i = 1,\ldots,m\}$ the cost function measures how well a given predictor or hypothesis h_w predicts the class of a sample.

- $\mathrm{J}(w) = -\sum_{\mathrm{i}} (y^{(i)} \log(\mathrm{h}_w (x^{(i)})) + (1 - y^{(i)}) \log(1 - \mathrm{h}_w (x^{(i)})))$.

- Only one term in the cost function is non-zero for each training example depending on whether the label $y^{(i)}$ is 0 or 1.

- When $y^{(i)} = 1$ minimizing the cost function means we need to make $\mathrm{h}_w(x^{(i)})$ large, and when $y^{(i)} = 0$ we want to make $1- \mathrm{h}_w(x^{(i)})$ large.

Classification

- We can learn to classify our training data by minimizing J(w) to find the best choice of w.

- We can classify a new test point as "1" or "0" by checking which of these two class labels is most probable: if P(y = 1|x) > P(y = 0|x) then we label the example as a "1", and "0" otherwise.

- This is the same as checking whether $h_w(x) > 0.5$.

- To minimize J(w): $\partial J(w)/\partial w\mathrm{j} = \sum_{\mathrm{i}} x^{(\mathrm{i})} j\ (h_w(x^{(\mathrm{i})}) - y^{(i)})$
- The entire gradient can be expressed as:
  $\nabla_w J(w) = \sum_{\mathrm{i}} x^{(\mathrm{i})}\ (h_w(x^{(\mathrm{i})}) - y^{(\mathrm{i})})$ where $h_{\mathrm{w}}(\mathrm{x}) = \sigma(\mathbf{w}^\top \mathbf{x})$

Hypothesis test

- A *hypothesis test* is a technique for validate or invalidate the claim about a population using sample data.
- This claim that's on trial, is called the ***null hypothesis***- the hypothesis assumes there is no relationship between the experimental variable(s) and the observed results
- The *alternative hypothesis* is the one you would believe if the null hypothesis is concluded to be untrue.
- The evidence in the trial is your data and the statistics that go along with it.

Test a hypothesis about a population, using test statistic to decide whether to reject the null hypothesis.

P value test

- Decision is based on a number, called a *p*-value.

- The purpose of finding a *p*-value is basically to determine whether the observed results differ from the expected results to such a degree that the "null hypothesis” - is unlikely enough to reject.

- The *p*-value is a number between 0 and 1 and interpreted as a significance level.

Choose a Significance Level

- Significance level is a measure of how certain we want to be about our results - low significance values correspond to a low probability that the experimental results happened by chance, and vice versa.

- Scientists usually set the significance value for their experiments at 0.05, or 5 percent.

- This means that experimental results that meet this significance level have, at most, a 5% chance of being reproduced in a random sampling process.

- Use a chi square distribution table to approximate your *p*-value.

- Chi square is a numerical value that measures the difference between an experiment's *expected* and *observed* values.

Softmax Regression

- Softmax regression is a generalization of logistic regression.

- Handle multiple classes unlike logistic regression where the labels are binary: $y^{(i)} \in \{0,1\}$.

- Softmax regression handles $y^{(i)} \in \{1,\ldots,K\}$ where K is the number of classes.

- For example, in the MNIST digit recognition task, we would have K=10 different classes.

- Given a test input **x**, we estimate the probability $P(\mathbf{y} = k|\mathbf{x})$ for each value of $k = 1,\ldots,K$.

Softmax Regression

- Output a K-dimensional vector (whose elements sum to 1) giving us our K estimated probabilities.

- $h_w(x) = P(y = 1|x,w)P(y = 2|x,w) \ldots\ldots. P(y = K|x,w)$

- In *softmax regression* (SMR), we replace the sigmoid logistic function by the so-called *softmax function* φ:

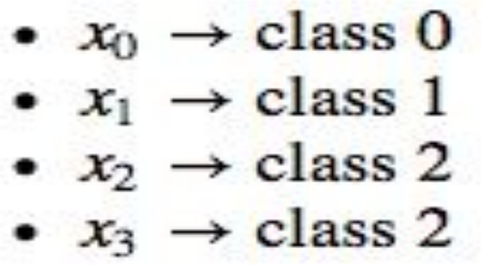
$$P(y = j \mid z^{(i)}) = \phi_{softmax}(z^{(i)}) = \frac{e^{z^{(i)}}}{\sum_{j=0}^{k} e^{z_k^{(i)}}},$$

Softmax Regression

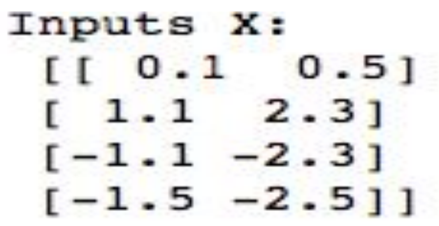
- Output a K-dimensional vector (whose elements sum to 1) giving us our K estimated probabilities.
- $h_w(x) = P(y = 1|x,w)P(y = 2|x,w)\ldots\ldots.P(y = K|x,w)$
- In *softmax regression* (SMR), we replace the sigmoid logistic function by the so-called *softmax function* φ:

$$P(y = j \mid z^{(i)}) = \phi_{softmax}(z^{(i)}) = \frac{e^{z^{(i)}}}{\sum_{j=0}^{k} e^{z_k^{(i)}}},$$

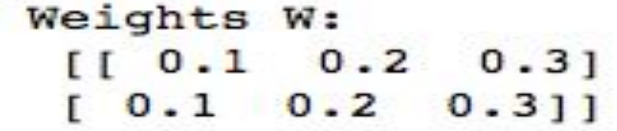
- Let's assume we have a training set consisting of 4 samples from 3 different classes (0, 1, and 2).



- $x_0 \rightarrow$ class 0
- $x_1 \rightarrow$ class 1
- $x_2 \rightarrow$ class 2
- $x_3 \rightarrow$ class 2

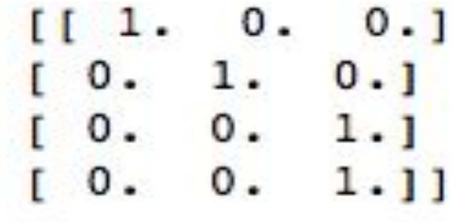


```
Inputs X:
 [[ 0.1   0.5]
 [ 1.1   2.3]
 [-1.1  -2.3]
 [-1.5  -2.5]]
```

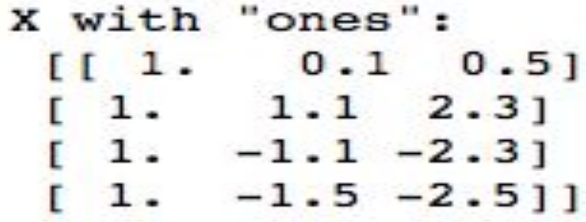


```
Weights W:
 [[ 0.1   0.2   0.3]
 [ 0.1   0.2   0.3]]
```

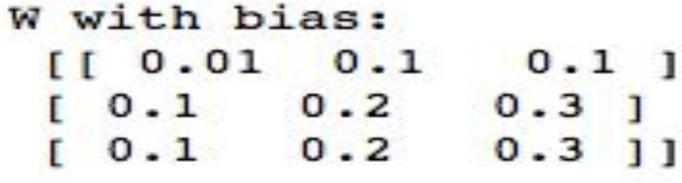
Output is encoded as a three dimensional vector:



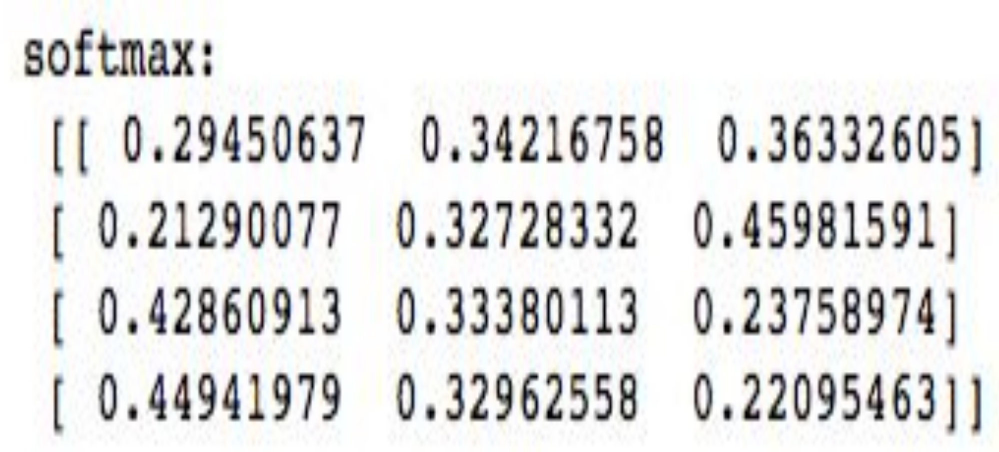
```
[[ 1.  0.  0.]
[ 0.  1.  0.]
[ 0.  0.  1.]
[ 0.  0.  1.]]
```



```
X with "ones":
 [[ 1.    0.1   0.5]
 [ 1.    1.1   2.3]
 [ 1.   -1.1  -2.3]
 [ 1.   -1.5  -2.5]]
```



```
W with bias:
 [[ 0.01   0.1    0.1  ]
 [ 0.1    0.2    0.3  ]
 [ 0.1    0.2    0.3  ]]
```



```
softmax:
 [[ 0.29450637  0.34216758  0.36332605]
 [ 0.21290077  0.32728332  0.45981591]
 [ 0.42860913  0.33380113  0.23758974]
 [ 0.44941979  0.32962558  0.22095463]]
```

Principal Components Analysis

- Background mathematics to understand the process of Principal Components Analysis covers *standard deviation*, *covariance*, *eigenvectors* and *eigenvalues*.

- Statistical analysis is based on the idea that we have big set of data, and want to analyse that set in terms of the relationships between the individual points in that data set.

- The mean is not enough to tell the characteristics of samples of data.
- Ex. [0 8 12 20] and [8 9 11 12] : same mean value (10) but the spread of data is different.

Principal Components Analysis

- The Standard Deviation (SD) of a data set is a measure of how spread out the data is.

- The average distance from the mean of the data set to a point is the measure of SD.

$$SD\ (\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(X_i \ - X_{mean})^2}{(n-1)}}$$

- Standard deviation of data set [0 8 12 20] is more than [8 9 11 12] due to the fact that the earlier data is much more spread out from the respective mean.

- Ex. [10 10 10 10]; A mean of 10, but its standard deviation is 0

Variance

- Variance is another measure of the spread of data in a data set.

$$variance\ (\sigma^2) = \frac{\sum_{i=1}^{n}(X_i - X_{mean})^2}{(n-1)}$$

Standard deviation and variance only operate on 1 dimension, so we analyse each dimension of the data set *independently* of the other dimensions.

For more than one dimensional data set, we see if there is any relationship between the dimensions using Covariance measure

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - X_{mean})\,(Y_i - Y_{mean})}{n-1}$$

The covariance matrix is a *p* × *p* symmetric matrix (where *p* is the dimension)

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Covariance Matrix for 3-Dimensional Data

The diagonal elements represent the variances of each variable. Sign of each element is important.

if the covariance is zero, it indicates that the two dimensions are independent of each other.

Linear Transformation

- **Av** = **b**; **A** is a known matrix, **v** is a unknown vector and **b** is another known vector.
- The matrix **A** performs a linear transformation on the vector **v** and maps it to vector **b**
- New vector **b** is projection of **v**
- An *eigenvector* of a square matrix **A** is a non-zero vector **u** such that multiplication by **A** alters only the scale of **u**: **Au** = λ**u**.
- Eigenvectors only changes its length not direction.
- Eigenvectors are representative of the matrix and find characteristics by decomposing into eigenvectors.

Eigenvectors and Eigenvalues of a matrix

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 11 \\ 5 \end{pmatrix}$$

- $[11 \quad 5]^T$ is not an integer multiple of the original vector.

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

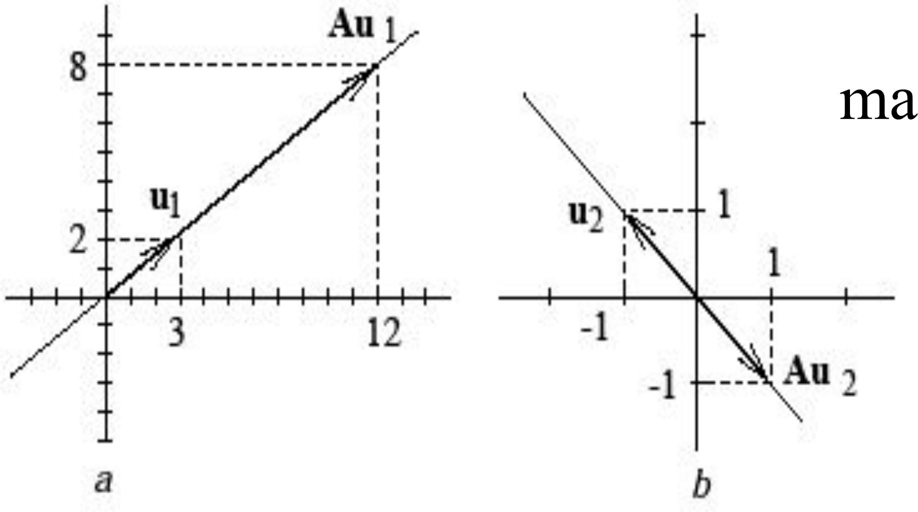
- The vector $[3 \quad 2]^T$ represents an arrow pointing from the origin (0, 0) to the point (3, 2). After matrix multiplication, we obtain the vector 4 times of the original.
- The square matrix is a **transformation matrix**.
- If we multiply this matrix with a vector, the answer is another vector that is transformed of its original .

Property of Eigenvector

- Eigenvectors can only be found for square matrices but not every square matrix has eigenvectors.
- Given an *n*×*n* matrix that does have *n* eigenvectors.
- A vector if scaled before multiplication generate the same multiple of it as a result, not changing direction.

$$2\begin{pmatrix}3\\2\end{pmatrix}=\begin{pmatrix}6\\4\end{pmatrix} \qquad \begin{pmatrix}2 & 3\\2 & 1\end{pmatrix}\times\begin{pmatrix}6\\4\end{pmatrix}=\begin{pmatrix}24\\16\end{pmatrix}=4\times\begin{pmatrix}6\\4\end{pmatrix}$$

Eigenvector



matrix **A** = $\begin{matrix} 2 & 3 \\ 2 & 1 \end{matrix}$

$\mathbf{u}_1 = [3 \quad 2]^T$, $\mathbf{u}_2 = [-1 \quad 1]^T$

For most applications we normalize the eigenvector with unit length.

- The trace operator gives the sum of all of the diagonal entries of a matrix $Tr(A) = \Sigma_i A_{ii}$ and $Tr(A) = \sum_i \lambda_i$

A is symmetric

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \qquad det\begin{pmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{pmatrix} = 0$$

$\lambda_1 = 4$ and $\lambda_2 = 2$

Eigenvector corresponding to $\lambda_1 = 4$ is $\boldsymbol{x} = [1 \;\; 1]^T$

Eigenvector corresponding to $\lambda_1 = 2$ is $\boldsymbol{y} = [-1 \;\; 1]^T$

- A vector ***x*** and a vector ***y*** are *orthogonal* to each other if $\boldsymbol{x}^T\boldsymbol{y} = 0$
- We can express the data in terms of these perpendicular eigenvectors, instead of the *x* and *y* axis

Eigenvalue

- The amount by which the original vector was scaled after multiplication by the square matrix, called *eigenvalue.*; $\mathbf{Au} = \lambda\mathbf{u}$
- $\mathbf{Au} - \lambda\mathbf{u} = 0$; $\mathbf{Au}-\lambda\mathbf{Iu} = 0$; $(\mathbf{A}-\lambda\mathbf{I})\mathbf{u} = 0$, $\mathbf{I}$ is the Identity matrix
- If $\mathbf{u}$ is non-zero, $(\mathbf{A}-\lambda\mathbf{I})\mathbf{u} = 0$ has a nonzero solution to $\mathbf{u}$ if and only if $(\mathbf{A}-\lambda\mathbf{I})$ has a non-empty null space, which is only the case if $(\mathbf{A}-\lambda\mathbf{I})$ is singular; i.e. $|(\mathbf{A}-\lambda\mathbf{I})| = \mathbf{0}$
- *Eigenvalue* equation is called the Characteristic equation of **A**, and is an n^{th} order polynomial in λ with *n* roots.
- These roots are called the *eigenvalues* of **A**, may be repeated.
- *Eigenvectors* and *eigenvalues* always come in pairs.

Matrix Decomposition

- Matrix decompositions are a useful tool for reducing a matrix to their constituent parts in order to simplify a range of more complex operations.
- Eigen decomposition decomposes a square matrix into eigenvectors and eigenvalues.
- This decomposition plays a role in Principal Component Analysis method or PCA.
- We can reverse the process and reconstruct the original matrix given only the eigenvectors and eigenvalues.

Eigendecomposition

- Suppose a matrix $\mathbf{A}$ has n linearly independent *eigenvectors*, $\{\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)}\}$, with corresponding *eigenvalues* $\{\lambda_1, \ldots, \lambda_n\}$.
- We concatenate the eigenvectors to form a matrix $\mathbf{U}$ with one *eigenvector* per column: $\mathbf{U} = [\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)}]$.
- We concatenate the *eigenvalues* to form a vector
  $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]^T$
- The *eigendecomposition* of $\mathbf{A}$ is then given by
  $\mathbf{A} = \mathbf{U}\,\mathrm{diag}(\boldsymbol{\lambda})\mathbf{U}^{-1}$
- Not every matrix can be decomposed into eigenvalues and eigenvectors.

Principal component analysis (PCA)

Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset.

- PCA method is used for dimensionality reduction of large data sets, still contains most of the information of the large set.
- A Multivariate Analysis problem useful for the dataset with substantial number of correlated variables. The new set of variables, called principal components (PCs), are uncorrelated.
- However, they are not just a one-to-one transformation, so inverse transformations are not possible.

Principal components

- Principal components are eigenvectors of the data's covariance matrix.

- The principal components are computed by eigendecomposition of the data covariance matrix.

- The steps of PCA are standardization, covariance, eigenvectors and eigenvalues .

- The first step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.

Standrdization

- Subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{value - mean}{standard\ deviation}$$

- This produces a data set whose mean is zero and scaled down by SD

Covariance Matrix

- Understand how the variables of the input data set are varying from the mean with respect to each other.

- To identify the correlation between the variables, we compute the covariance matrix.

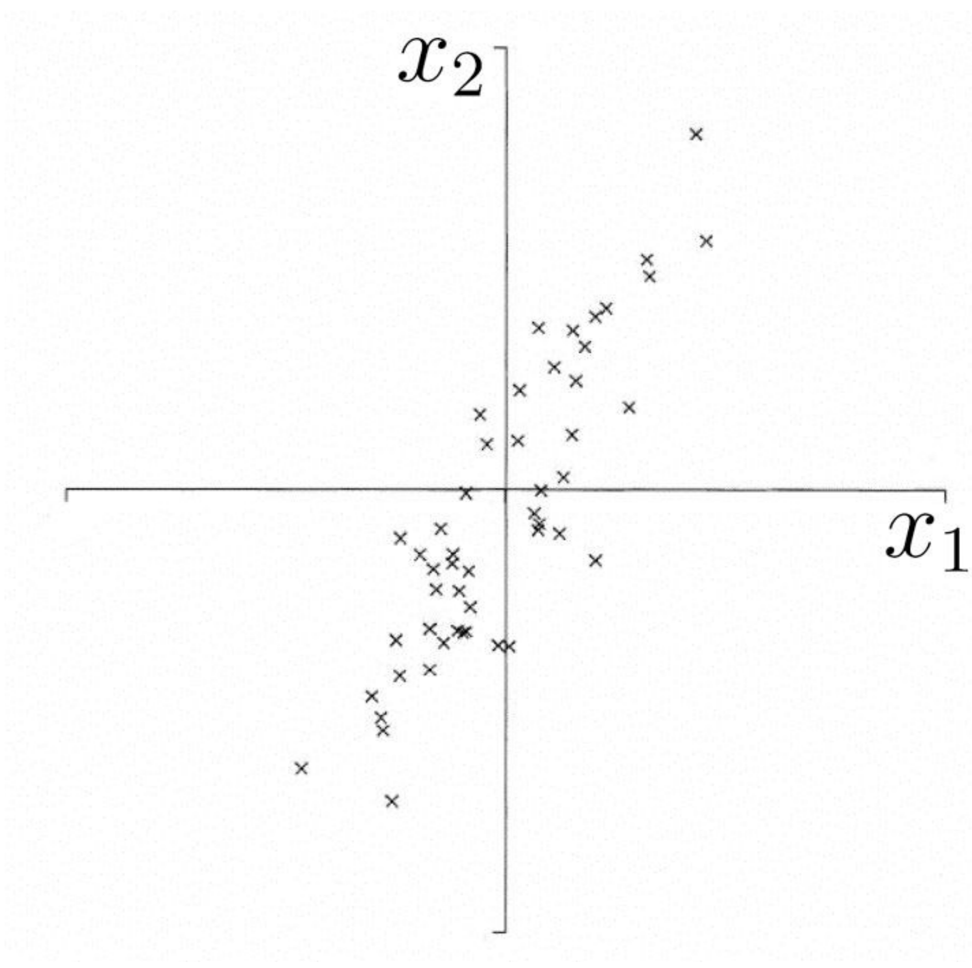
Eigenvectors and Eigenvalues of the Covariance Matrix

- Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables.

- These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated .

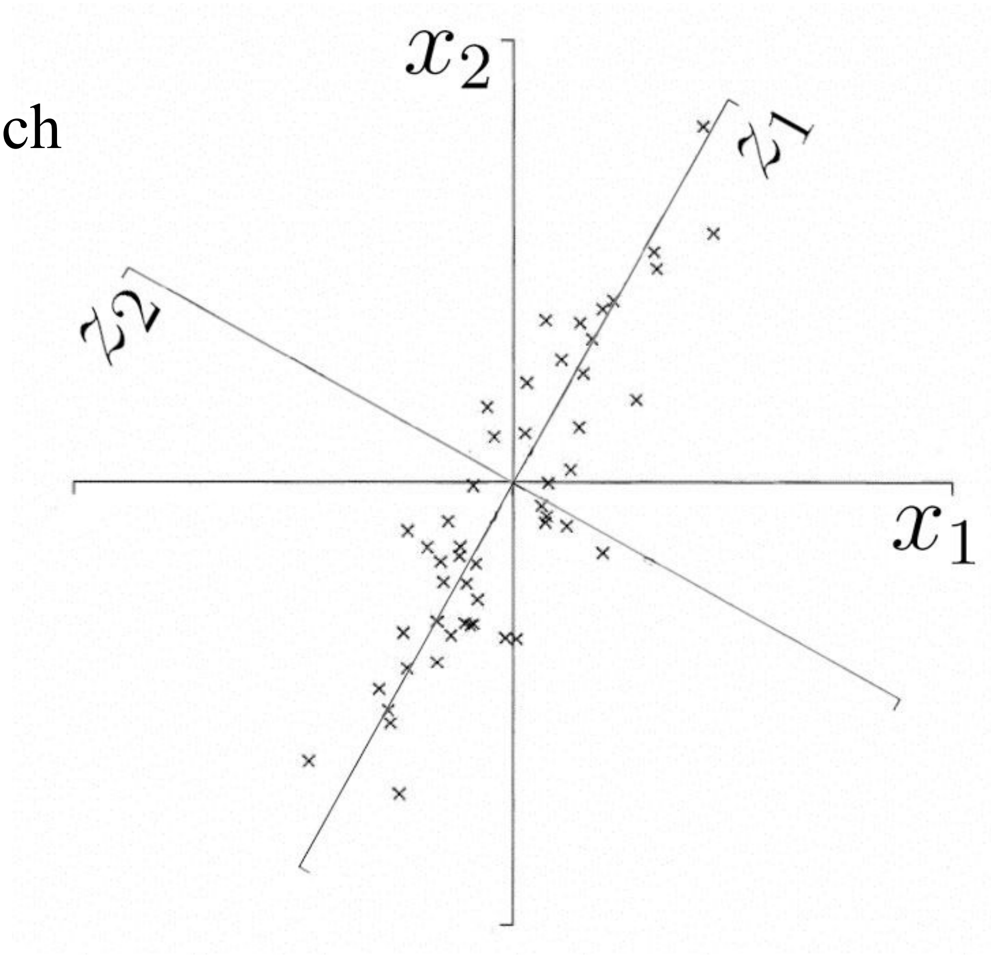
- Most of the information within the initial variables is compressed into the first components and so on.

- the principal components are less interpretable and don’t have any real meaning since they are constructed as linear combinations of the initial variables.



A sample of *n observations in the 2-D space* $X = (x_1, x_2)$

A new coordinate system in which every point has a new value.



- The z_1 vector is the direction of a line that best fits the data, defined as one that minimizes the average squared <u>distance from the points to the line</u> (the variance of the projected data).

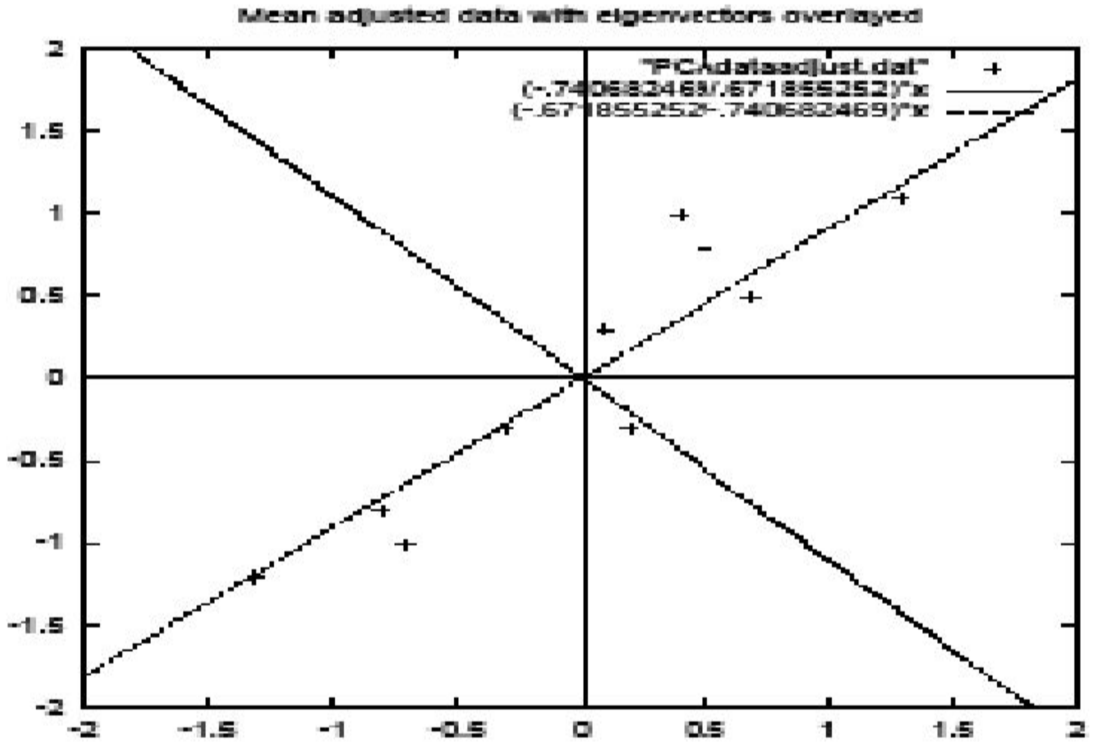
- The z_2 vector is perpendicular to the z_1 vector

$$cov = \begin{pmatrix} .616555 & .615444 \\ .615444 & .71555 \end{pmatrix}$$

$$eigenvalues = \begin{pmatrix} 0.0490 \\ 1.2840 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.7351 & -.6778 \\ .6778 & -.7351 \end{pmatrix}$$

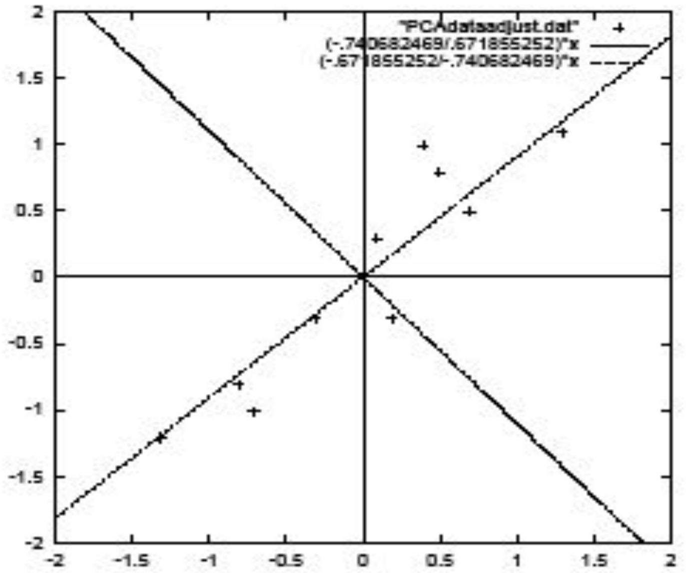
- eigenvectors are both *unit eigenvectors ie. Their* lengths are both 1.


Mean adjusted data with eigenvectors overlayed
"PCAdataadjust.dat" +
(-.740682469/.671855252)*x
(-.671855252/-.740682469)*x
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

Choosing components

- The eigenvector with the *highest eigenvalue is the principle component of the data set.*

- The eigenvector with the largest eigenvalue was the one that pointed down the middle of the data.



A plot of the normalised data (mean subtracted) with the eigenvectors of the covariance matrix overlayed on top.

Feature vector:
(eig1, eig2…eign)

PCA

- There are as many PCs as there are dimension of feature and we prioritize them for selection based on the variance.

- Finding eigenvectors of the covariance matrix is equivalent of finding principal components (PCs) to the variance of data.

- Let the vectors $\mathbf{e}_1$ through $\mathbf{e}_p$ *corresponding to eigenvalues* λ_1 *to* λ_p
- The elements for these eigenvectors are the coefficients of our principal components.

- The variance for the *i*-th PC is equal to the *i*-th eigenvalue.