Discriminant Analysis- Linear and Gaussian

Shaily jain · Follow
4 min read · Apr 20, 2021

Now Logistic Regression and Multinomial Regression are called *Discriminant learning algorithms* which learn p(y|x) directly.

Naive bayes and linear/quadratic discriminant analysis are called *Generative learning algorithms* that try to model p(x|y) and p(y). They use Bayes rule to derive p(y|x).

Discriminant Analysis

Discriminant analysis seeks to model the distribution of X in each of the classes separately. Bayes theorem is used to flip the conditional probabilities to obtain P(Y|X). The approach can use a variety of distributions for each class. The techniques discussed will focus on normal distributions

Linear Discriminant Analysis:

With linear discriminant analysis, there is an assumption that the covariance matrices Σ are the same for all response groups.

$$The\ model\ is$$

$$y \sim Bernoulli(\phi)$$

$$x|y = 0 \sim \mathcal{N}(\mu_0, \Sigma)$$

$$x|y = 1 \sim \mathcal{N}(\mu_1, \Sigma)$$

For p(no. of independent variables)= 1:

Recall the pdf for the Gaussian distribution:

$$f_k(x) = \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu_k}{\sigma_k})^2}$$

Then

$$P(Y=k|X=x) = \frac{\pi_k \frac{1}{\sigma_k\sqrt{2\pi}} exp\left(-\frac{1}{2}(\frac{x-\mu_k}{\sigma})^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}(\frac{x-\mu_l}{\sigma})^2\right)}$$

where πk=P(Y=k).

Simplify by taking logs and simplifying

$$\log(P(Y=k|X=x)) = \frac{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \log(\pi_k) - \frac{x^2 - 2x\mu_k + \mu_k^2}{2\sigma^2}}{\log\left(\sum_{l=1}^{K} \pi_l \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}(\frac{x-\mu_l}{\sigma})^2\right)\right)}$$

Since the objective is to maximize, remove all constants (terms that do not depend on kk) to obtain the discriminant score

$$\delta_k(x) = \log(P(Y=k|X=x)) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Assign xx to the class with the largest discriminant score.

For p > 1:

The pdf for the multivariate Gaussian distribution:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

The discriminant function is

See the function is linear in x

This method assumes that the covariance matrix Σ is the same for each class.

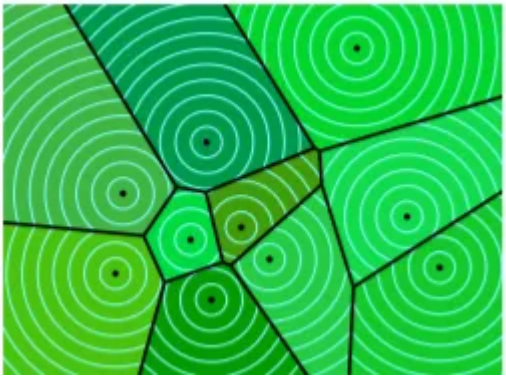
Estimate the model parameters using the training data.

$$\hat{\pi}_k = \frac{n_k}{n}$$
$$\hat{\mu}_k = \frac{1}{n_k}\Sigma_i x_i * I(y_i = k)$$
$$\Sigma = \frac{1}{n-k}\sum^K \sum_{i=1}^n (x-\mu_k)(x-\mu-K)^T \text{ where } k \text{ is the number of classes}$$

Compute the class posterior probabilities with the discriminant function

$$P(Y = k|X = x) = \frac{e^{\hat{\delta}_k}}{\sum_{l=1}^K e^{\hat{\delta}_l}}$$

Decision boundaries are pictorially represented like



[When you have many classes, their LDA decision boundaries form a classical Voronoi diagram if the priors ↑C are equal. All the Gaussians have the same width.]

Note: That the decision function

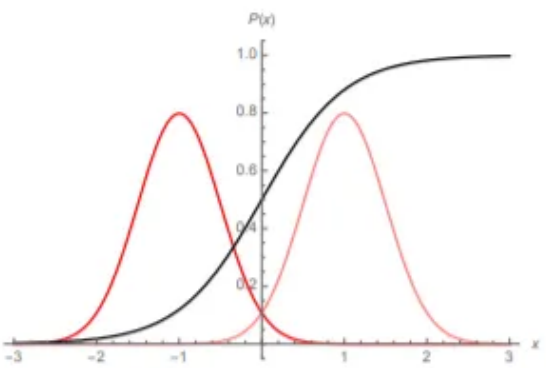
$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k)$$

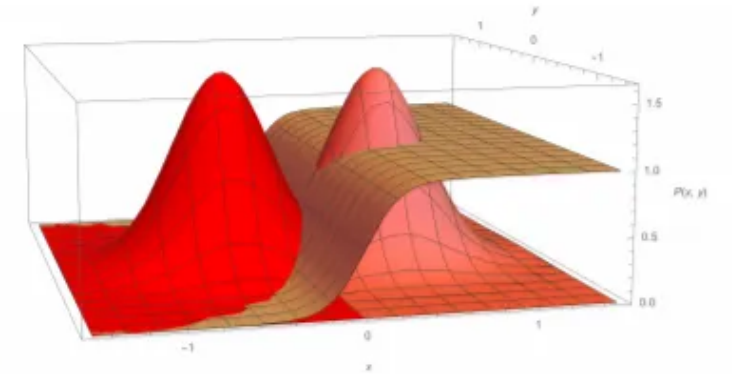
talked earlier also

in case of two cases, when subtracted for two classes give

$$\underbrace{\frac{(\mu_C - \mu_D) \cdot x}{\sigma^2}}_{w \cdot x} \underbrace{- \frac{\|\mu_C\|^2 - \|\mu_D\|^2}{2\sigma^2} + \ln \pi_C - \ln \pi_D}_{+a}$$

gives decision boundary when w.x +a = 0 .[The effect of "w · x + a" is to scale and translate the logistic fn in x-space. It's a linear transformation.]





The translation of posterior probabilities to logistic/ sigmoid function is

$$P(Y = C|X = x) = \frac{e^{Q_C(x)}}{e^{Q_C(x)} + e^{Q_D(x)}} = \frac{1}{1 + e^{Q_D(x) - Q_C(x)}}$$
$$= s(Q_C(x) - Q_D(x)), \quad \text{where}$$

$$s(\gamma) = \frac{1}{1 + e^{-\gamma}} \quad \Leftarrow \text{logistic fn aka sigmoid fn}$$

Here Q is delta in our case

Quadratic Discriminant Analysis

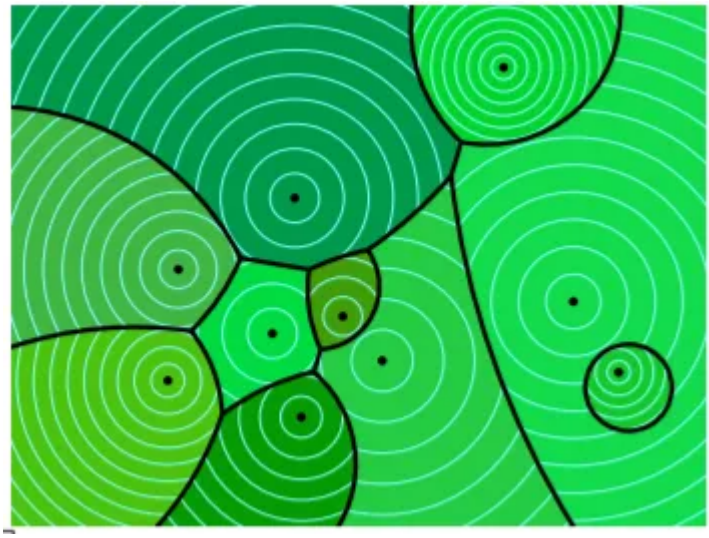
In quadratic discriminant analysis, do not make the assumption that the covariance matrix Σk is the same for each class.

This changes the discriminant function to

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x - \mu_k)^T\Sigma_k^{-1}(x - \mu_k) + \log(\pi_k)$$

See the quadratic expression in X here

The decision boundaries are pictorially represented as



f [The feature space gets partitioned into regions. In two or more dimensions, you typically wind up with multiple decision boundaries that adjoin each other at joints. It looks like a sort of Voronoi diagram. In fact, it's a special kind of Voronoi diagram called a multiplicatively, additively weighted Voronoi diagram.]

Please Notes we might get coefficients are an output while implementing LDA, which simply means that coeff*variable value for each observation gives the decision rule, if there are only two class labels then,?

Please check NAIVE BAYES for generative algorithm for classification

Logistic Regression vs. Discriminant Analysis vs. Naive Bayes

Best to use Logistic Regression:

- More robust to deviations from modeling assumptions (non-Gaussian features)

Best to use Discriminant Analysis:

- When the assumption that the features are Gaussian can be made
- More efficient than logistic regression when the assumptions are correct
- Works better than logistic regression when data is well-separated
- Popular for multinomial responses since it provides a low-dimensional view of data

Best to use Naive Bayes:

- Can make the assumption that features are independent (conditional on response)
- Despite strong assumptions, works well on many problems

Resources:

https://people.eecs.berkeley.edu/~jrs/189/lec/07.pdf

http://jennguyen1.github.io/nhuyhoa/statistics/Discriminant-Analysis-Naive-Bayes.html

Stats | Statistics | Algorithms | Machine Learning | Data Science Training

Written by Shaily jain

37 Followers

Problem Solver, Data Science, Actuarial Science, Knowledge Sharer, Hardcore Googler

More from Shaily jain

Shaily jain

Impurity Measures

Let’s start with what they do and why we need them.

5 min read · Apr 29, 2021

5

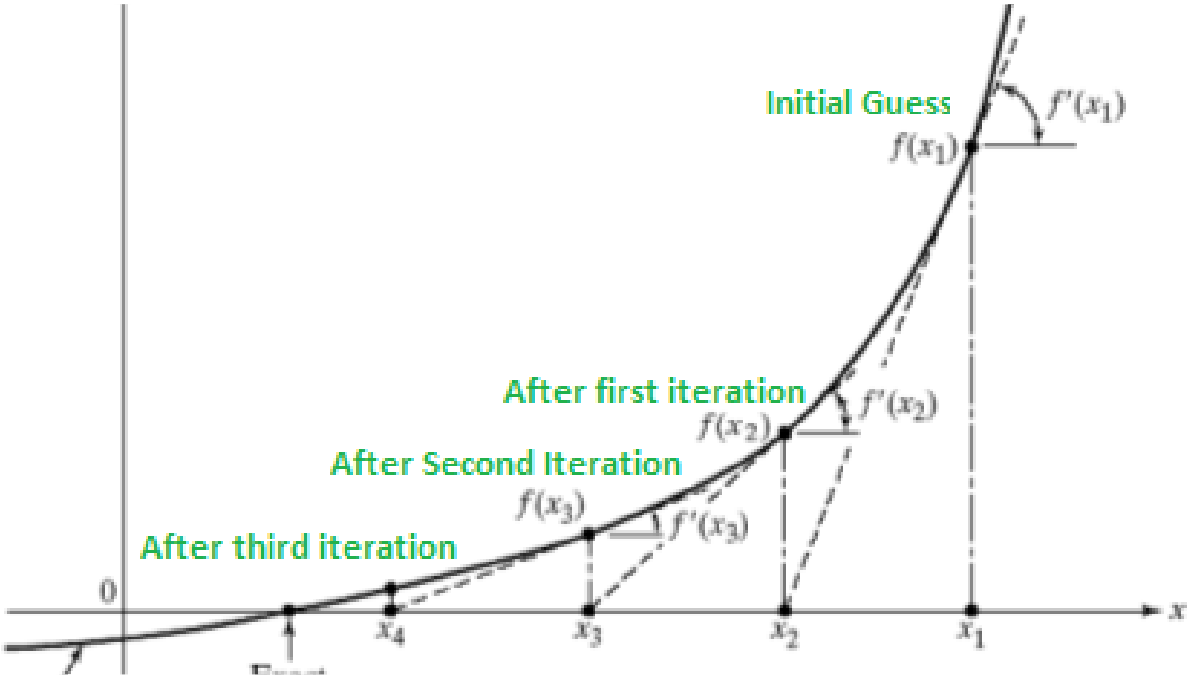
Shaily jain

Limitations, Assumptions Watch-Outs of Principal Component Analysis

Hey!, I know enough of PCA, but why not?

5 min read · May 15, 2021

1



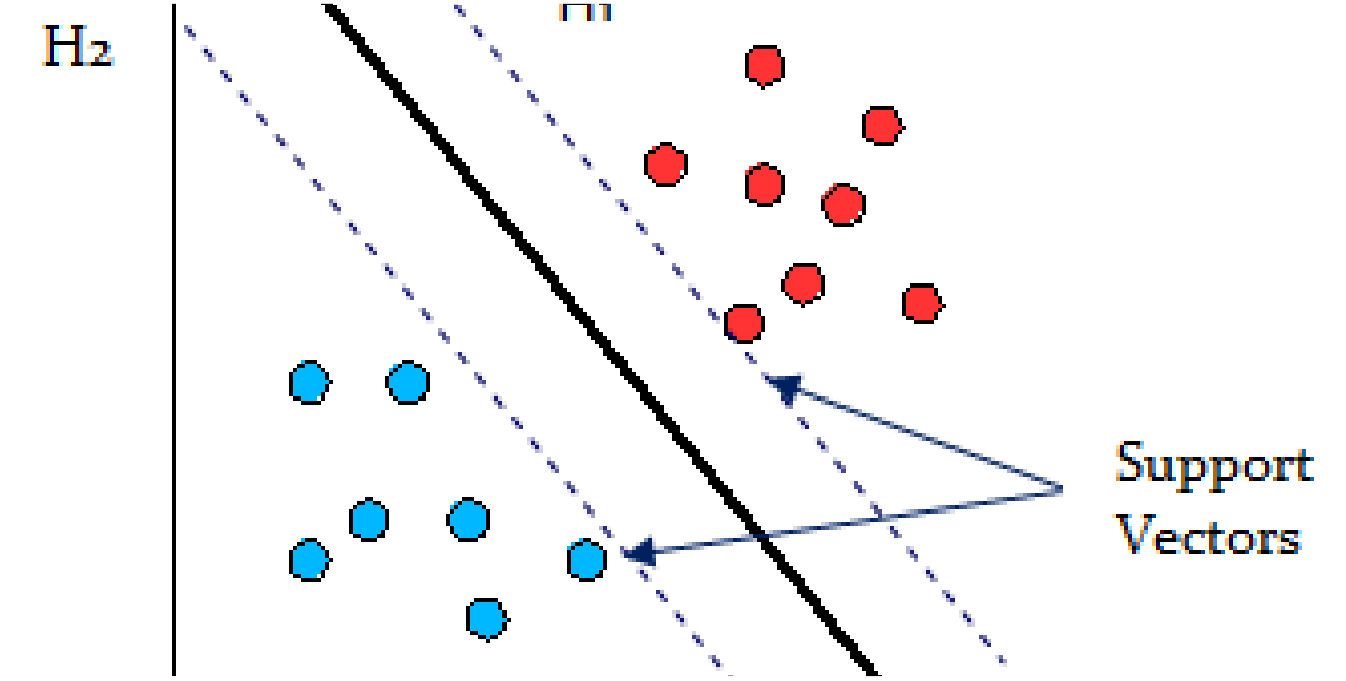
Shaily jain

Newton's Method for Logistic Regression

Optimisation Technique

3 min read · Sep 19, 2020

5



Shaily jain

Support Vector Machine

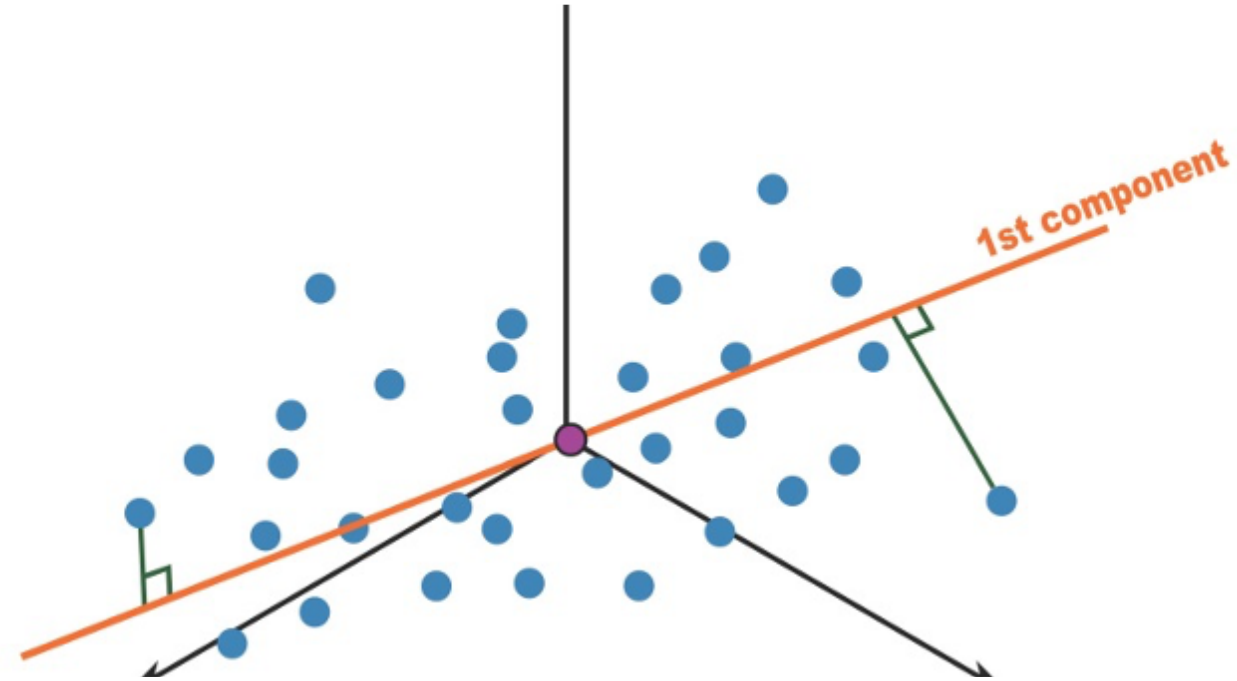
Maximal Margin Classifier

5 min read · Feb 26, 2021

3

See all from Shaily jain

Recommended from Medium



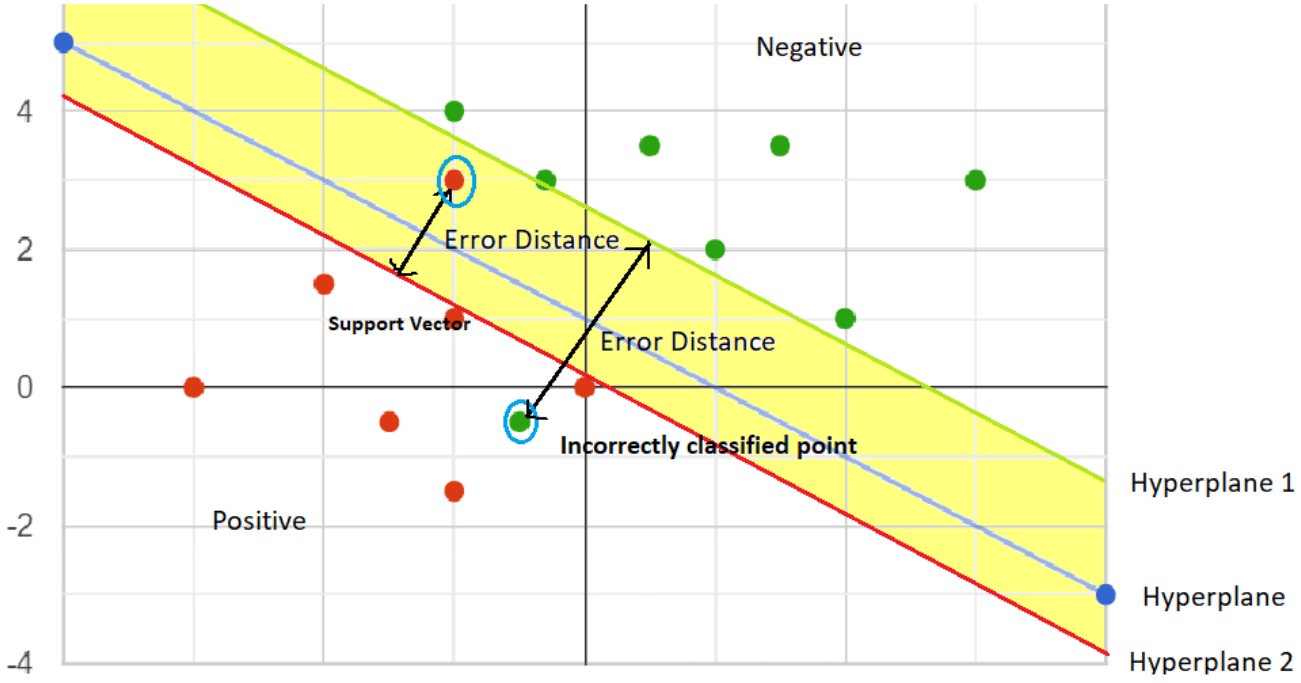
Huda Swati

Understanding Principal Component Analysis (PCA)

What is PCA?

8 min read · Sep 25, 2023

78



Kasun Dissanayake in Towards Dev

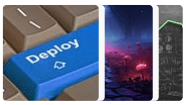
Machine Learning Algorithms(16)—Support Vector Machine(SVM)

This article, delves into the topic of Support Vector Machines(SVM) in Machine Learning, covering the different types of SVM algorithms...

12 min read · Feb 2

287 1

Lists



Predictive Modeling w/ Python

20 stories · 894 saves

Practical Guides to Machine Learning

10 stories · 1037 saves

Natural Language Processing

1187 stories · 660 saves

The New Chatbots: ChatGPT, Bard, and Beyond

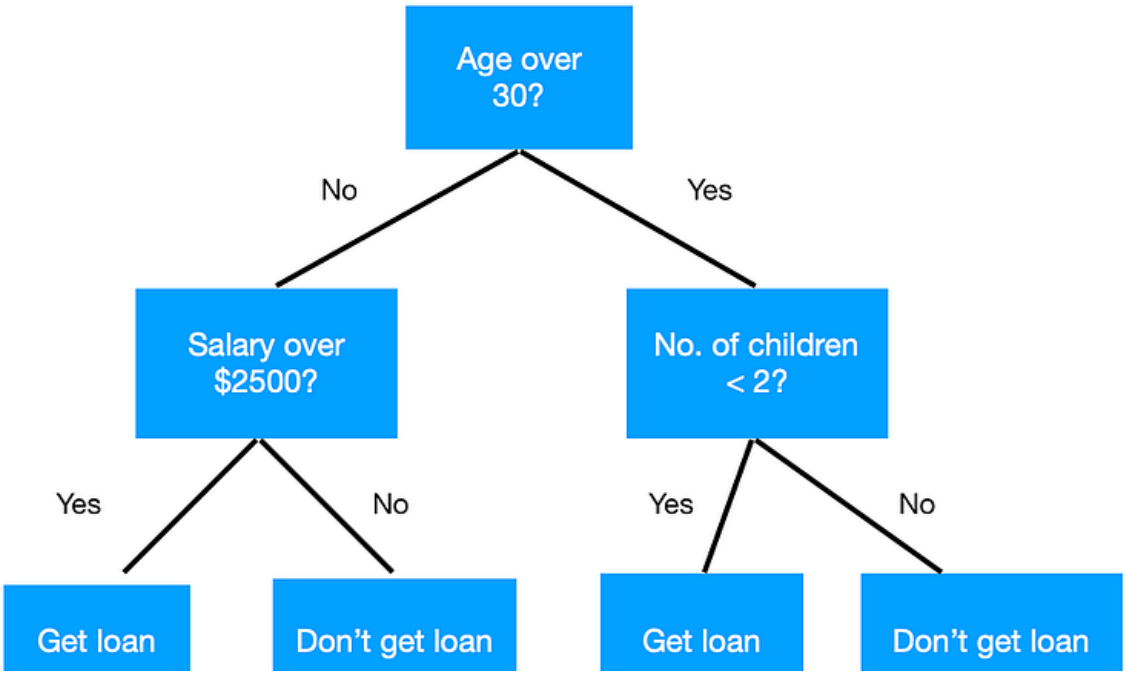
12 stories · 303 saves

David Mavrodiev

Understanding Probability Through Intuition

In this article, we'll explore the concept of probability in an intuitive manner, using real-life examples to demystify probability theory.

5 min read · 3 days ago



Kalim

Decision Tree

Introduction

5 min read · Aug 19, 2023

Setup	θ is unknown and fixed	θ is a random variable
What do we want?	$\hat{\theta}$ is the best estimate of θ (but fixed)	$p(\theta\|x,y)$ posterior distribution dictated by the data
What do we need?	Build a model using the data (x,y) *to* determine $\hat{\theta}$	Using the data determine $p(\theta)$ – prior probability
How we do it?	OLS: $\hat{\theta} = \frac{Cov(y,x_i)}{Var(x_i)}$ ML: $\mathcal{L}(x,\theta)$ likelihood function	$p(\theta\|x,y) \propto L(x,y\|\theta) * p(\theta)$ Find $p(\theta)$
Inference	H_0: $\theta = 0$ and H_a: $\theta \neq 0$ p-value is the probability that $\theta > \theta$	$p(\theta > \theta_c\|x,y)$ read from posterior probability

Roshmita Dey

Frequentist v/s Bayesian Statistics

Within the field of statistics, two major paradigms dominate the approach to inference: frequentist and Bayesian statistics. These...

6 min read · Jan 22

97 1

Parth Shah

Bayesian 101- Bayes Theorem

Hello, detectives!

2 min read · Aug 16, 2023

3 1

See more recommendations