Estatística II

Turma 941 | Santander Coders

Informações gerais da disciplina

Informações gerais da disciplina
Professor

- Thiago Tavares Magalhães

- Petrópolis - RJ

- Tecnologia da Informação (Faeterj Petrópolis)
- Mestrado em Modelagem Computacional (LNCC)
- Especialista em Otimização Matemática via Metaheurísticas
- Cientista de Dados Orange
- Cientista de Dados Meta|BRF (Sadia & Perdigão)
- Cientista de Dados Supersim
- Professor Let's Code by Ada

- Machine Learning com Python

- linkedin.com/in/thiagotm

Informações gerais da disciplina
Conteúdo, calendário e avaliação

➔ Conteúdo Abordado

- ➔ Introdução à teoria de aprendizagem de máquina
- ➔ Regressão linear simples
- ➔ Regressão linear múltipla
- ➔ Regressão logística
- ➔ Regularizações
- ➔ Generalizações com modelos lineares
- ➔ Análise de dados categóricos
- ➔ Redução de dimensionalidade

➔ Dias e Horários

- ➔ 22/05 a 12/06, segundas, quartas e sextas de 19:00h às 22:00h
- ➔ Atenção

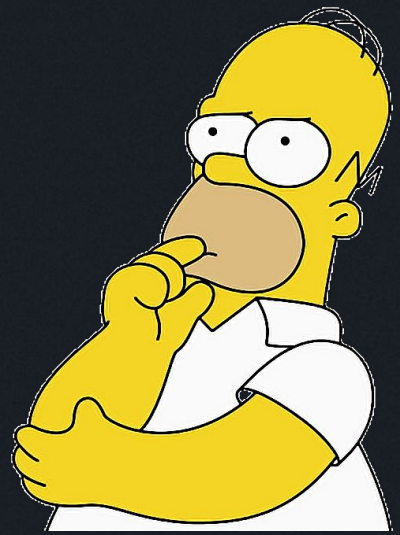
➔ Processo Avaliativo

- ➔ Avaliação por rubrica, com devolutiva e autoavaliação na aula do dia 12/06

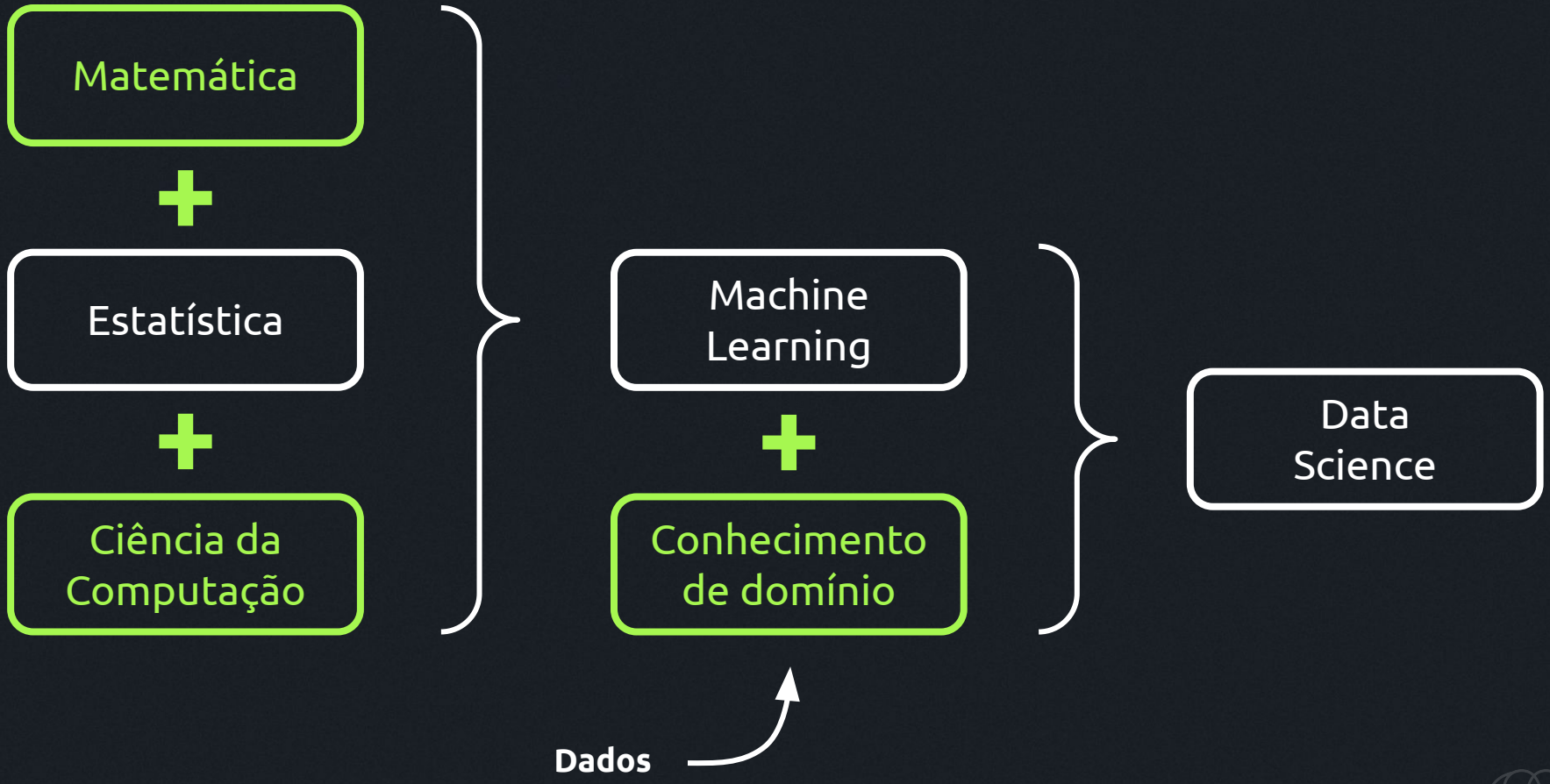
ada

Introdução à teoria de aprendizagem de máquina

Estamos no módulo de Estatística II ou de Machine Learning I?

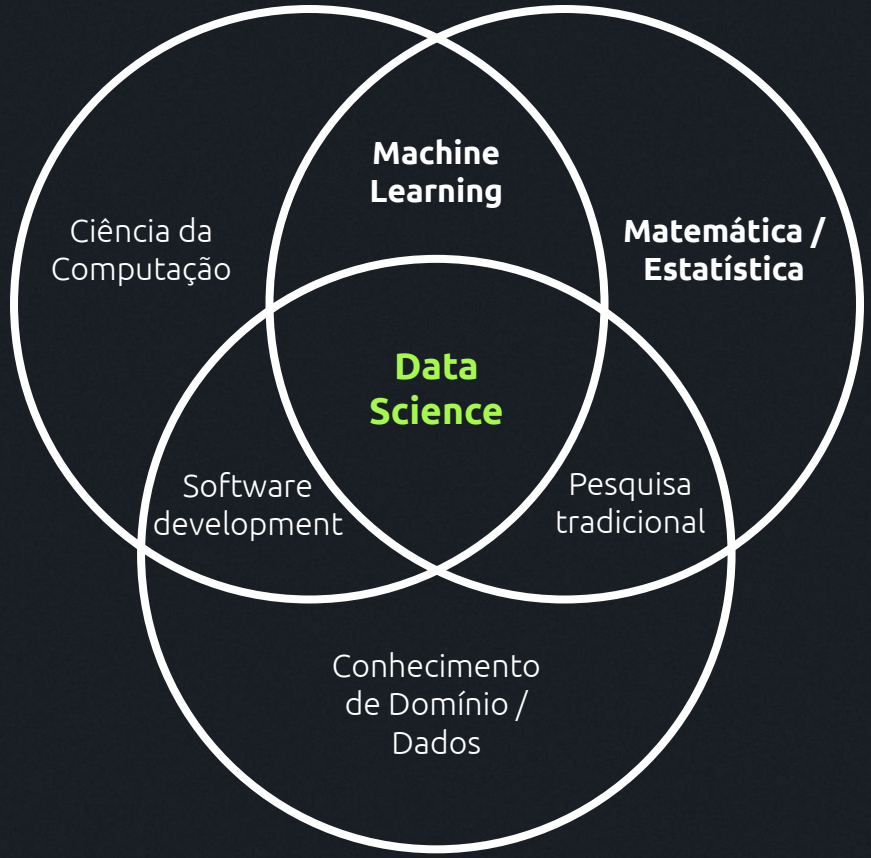


Estatística, Machine Learning e Data Science



Introdução à teoria de aprendizagem de máquina

Estatística, Machine Learning e Data Science



Conseguimos definir, então, cada uma das seguintes áreas?

Estatística

Inteligência Artificial

Aprendizagem de Máquina

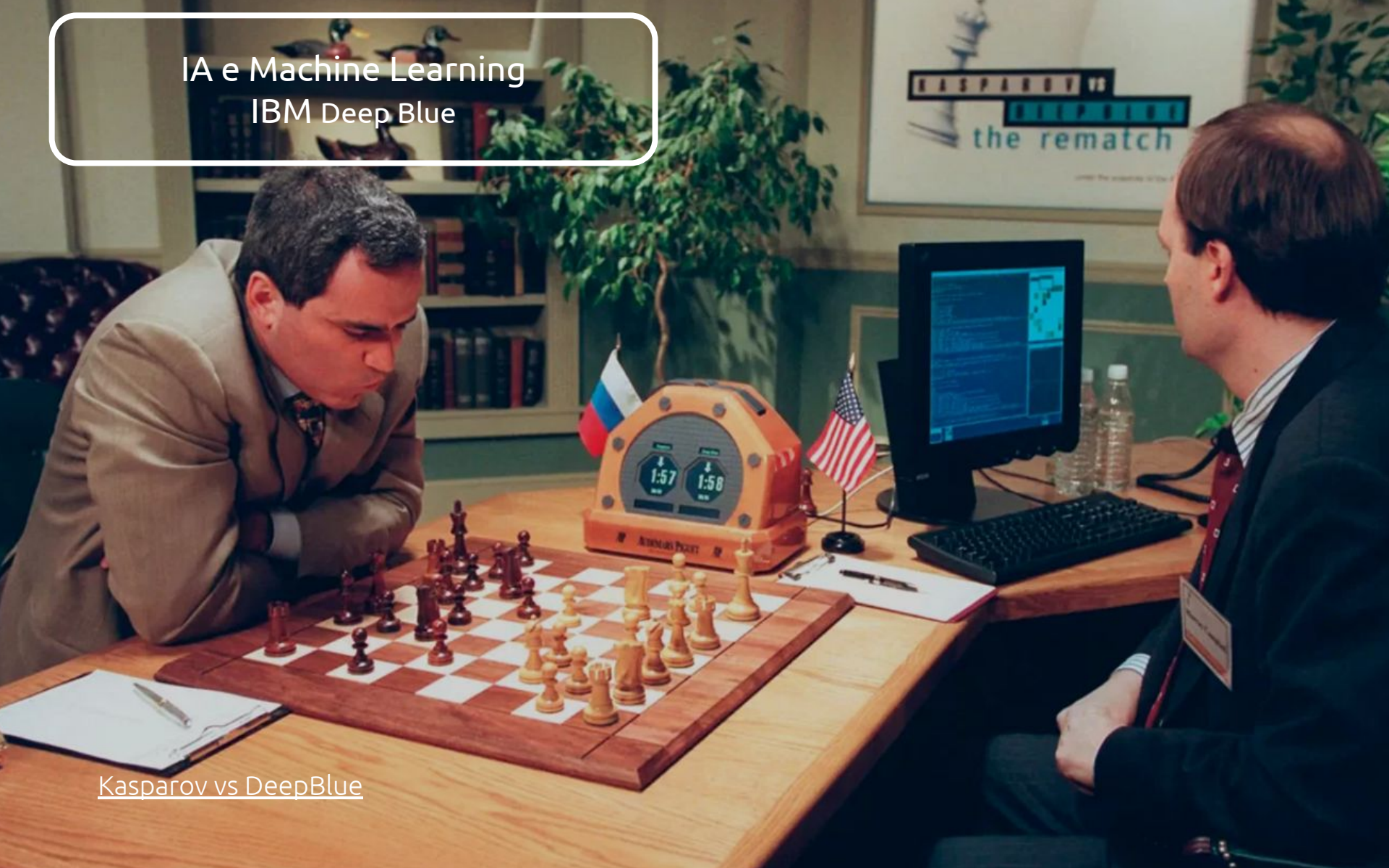
Estatística

Contas a pagar e relatórios financeiros de grandes empresas

Resultados Qualicorp

IA e Machine Learning

IBM Deep Blue



Kasparov vs DeepBlue

No contexto prático da indústria, como tem sido usado cada um destes termos?

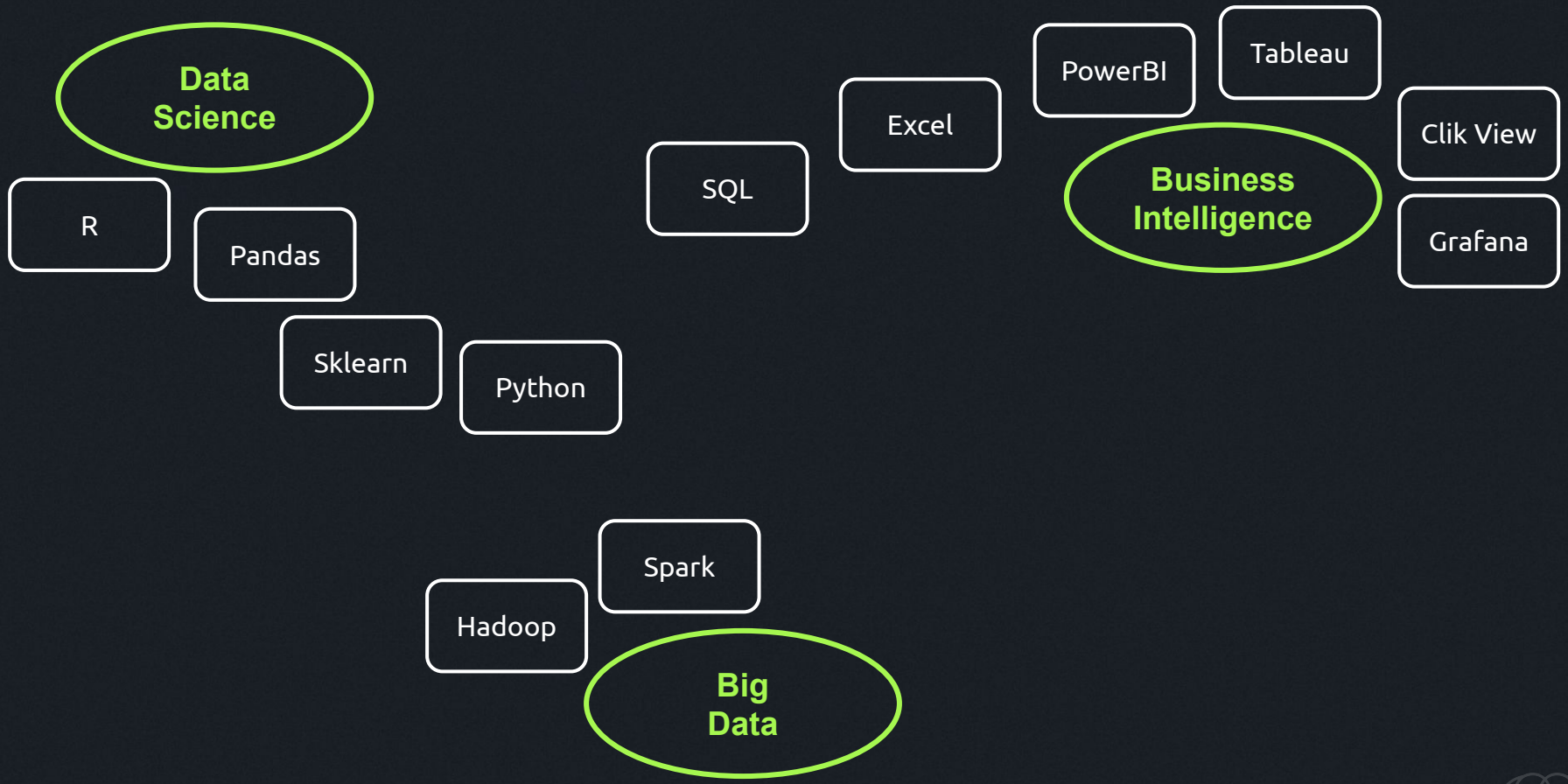
Business Intelligence

Big Data

Data Science

Introdução à teoria de aprendizagem de máquina

Terminologias



Conseguimos descrever e trazer exemplos de cada um dos principais tipos de análises de dados?

Análise Descritiva

Análise Prescritiva

Análise Preditiva

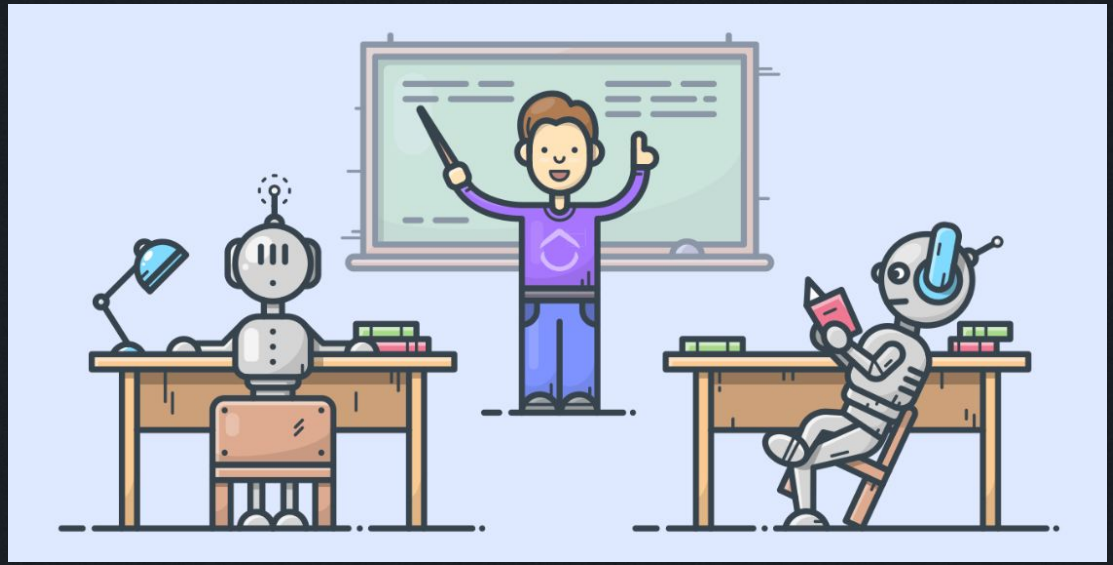
Terminologias

Análise Descritiva	Aumento da compreensão sobre os padrões, e tendências já observados nos dados
Análise Prescritiva	Prescrição de tomadas de decisões otimizadas com base nos dados
Análise Preditiva	Predição de comportamentos prováveis no futuro, baseadas em padrões presentes nos dados do passado e do presente

Terminologias

Entradas	Informações que descrevem cada uma das instâncias para as quais nós queremos uma resposta. Nós nos referimos aos diferentes tipos de informação como “features”.
Modelo de ML	Uma entidade matemático-computacional que através de um processo de treinamento se tornou capaz de mapear um conjunto de entradas em saídas.
Saídas	As respostas que se deseja obter para as entradas do problema. Podem ser probabilidades, classes, parâmetros de uma equação, recomendações, agrupamentos de objetos, ...

Conseguimos diferenciar os três principais tipos de aprendizagem de máquina?



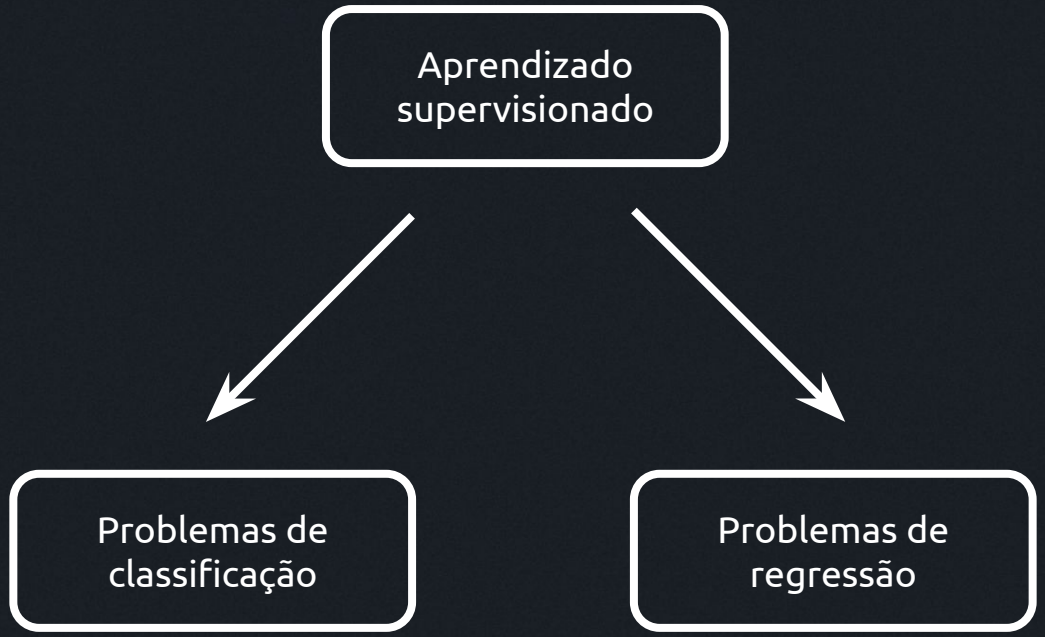
- Aprendizado supervisionado
- Aprendizado não-supervisionado
- Aprendizado por reforço

Introdução à teoria de aprendizagem de máquina

Terminologias

Aprendizado supervisionado	Aprendizado direcionado por uma classe que se deseja discriminar
Aprendizado não-supervisionado	Aprendizado sem a existência de uma classe, em que deseja separar o espaço de busca em grupos o máximo diferentes possível de acordo com algum critério
Aprendizado por reforço	Aprendizado baseado em um sistema de recompensas e penalizações incremental GP ant

Terminologias - problemas de aprendizado supervisionado



Aprendizado supervisionado: Fundamentos e métricas de classificação

Aprendizado supervisionado: classificação

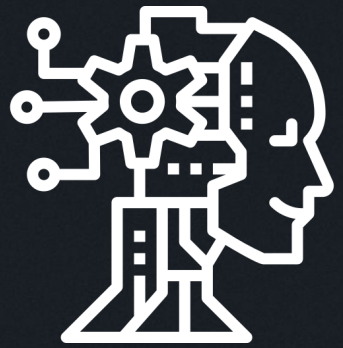
Definição

Modelo de ML

Aprendizado supervisionado: classificação

Definição

?



?

?

?

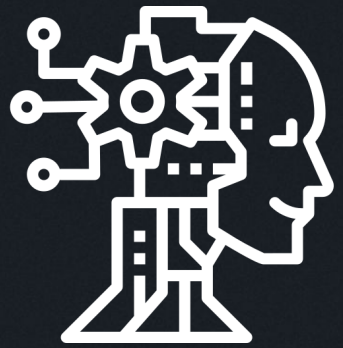
?

ada

Aprendizado supervisionado: classificação

Definição

Gato



Cachorro

Cachorro

Gato

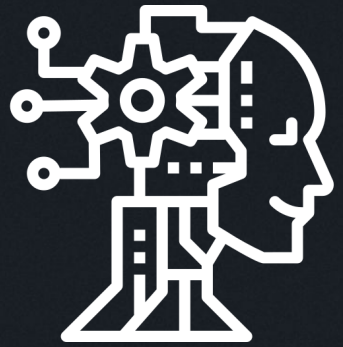
Gato

ada

Aprendizado supervisionado: classificação

Definição

0.83



0.47

0.12

0.53

0.79

ada

Definição

Variáveis atributos
(ou variáveis independentes, ou features)

Modelo de classificação
(ou classificador)

Variável resposta
(ou variável dependente, ou classe)

Variável nominal!

O que é pior:
um FN ou um FP ?

Aprendizado supervisionado: classificação

Métricas de avaliação

TP

True positive (verdadeiro positivo):
Registros que foram classificados como positivos pelo classificador e que eram, de fato, positivos

TN

True negative (verdadeiro negativo):
Registros que foram classificados como negativos pelo classificador e que eram, de fato, negativos

FP

False positive (falso positivo):
Registros que foram classificados como positivos pelo classificador mas que na realidade eram falsos

FN

False negative (falso negativo):
Registros que foram classificados como negativos pelo classificador mas que na realidade eram positivos

Accuracy

$$\frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Dentre todos os registros da base, qual porcentagem foi classificada corretamente?

Exemplo de conclusão
De todas as classificações que o modelo fez (TP+TN+FP+FN), 20% foram classificações corretas (TP+TN)

Nota
Pode trazer uma ideia muito ruim de desempenho quando em bases desbalanceadas. Supondo, por exemplo, uma base em que apenas 5% dos registros são atribuídos à classe positiva, se o modelo ignorar qualquer informação presente nos dados e simplesmente assumir que todos os registros devem ser classificados como N, já terá obtido uma acurácia de 95%.

Precision

$$\frac{(TP)}{(TP+FP)}$$

Dentre todos os registros classificados pelo modelo como positivos, qual porcentagem era de fato positiva?

Exemplo de conclusão
Dentre todos os pacientes para os quais o modelo reportou que o patógeno estava presente (TP+FP), 85% deles realmente tinham o patógeno presente (TP).

Nota
Mede a corretude das classificações positivas, mas sem levar em conta os registros que deveriam ter sido classificados desta forma e não foram (FN). Assim, um modelo pode registrar uma excelente precision porque opta por classificar um registro como P somente quando há uma evidência absolutamente clara que garanta essa classificação, errando, no entanto, para casos positivos cuja classificação correta seja ligeiramente menos óbvia.

Recall (ou sensibilidade)

$$\frac{(TP)}{(TP+FN)}$$

Dentre todos os registros na base que de fato são positivos, qual porcentagem o modelo foi capaz de classificar como tal?

Exemplo de conclusão
Considerando todos os pacientes que de fato apresentavam o patógeno presente (TP+FN), o modelo foi capaz de identificar como portadores 90% deles (TP).

Nota
Avalia quantos registros positivos da base foram devidamente detectados, mas se considerar a corretude na classificação dos negativos. Se, a despeito de qualquer informação presente nos dados, o modelo simplesmente escolher atribuir a classe positiva para todas as ocorrências, obviamente todos os casos positivos serão detectados e, portanto, o modelo alcançará o máximo recall.

Especificidade

$$\frac{(TN)}{(TN+FP)}$$

Dentre todos os registros na base que de fato são negativos, qual porcentagem o modelo foi capaz de classificar como tal?

Exemplo de conclusão
Considerando todos os pacientes que de fato já estavam livres do patógeno (TN+FP), o modelo foi capaz de identificar como curados 90% deles (TN).

Nota
É o exato oposto do recall, avaliando quantos registros negativos da base foram corretamente classificados como tal, mas sem considerar a corretude na classificação dos positivos. Da mesma forma, um modelo que opte simplesmente por atribuir a classe negativa a todo registro com o qual tenha contato, naturalmente identificará a totalidade dos negativos, atingindo a máxima especificidade.

F1 score

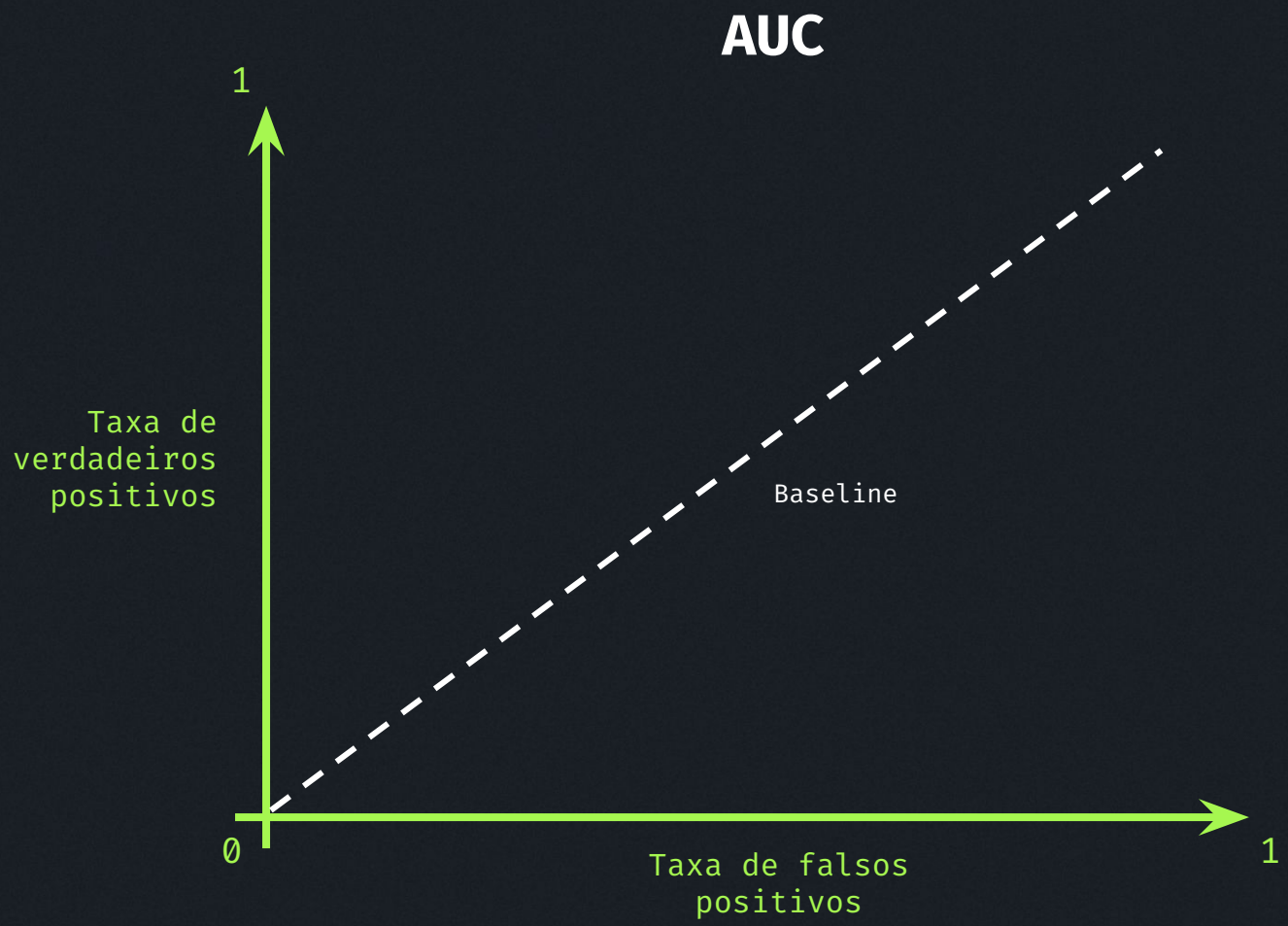
$$2 * \left(\frac{PRECISION * RECALL}{PRECISION + RECALL} \right)$$

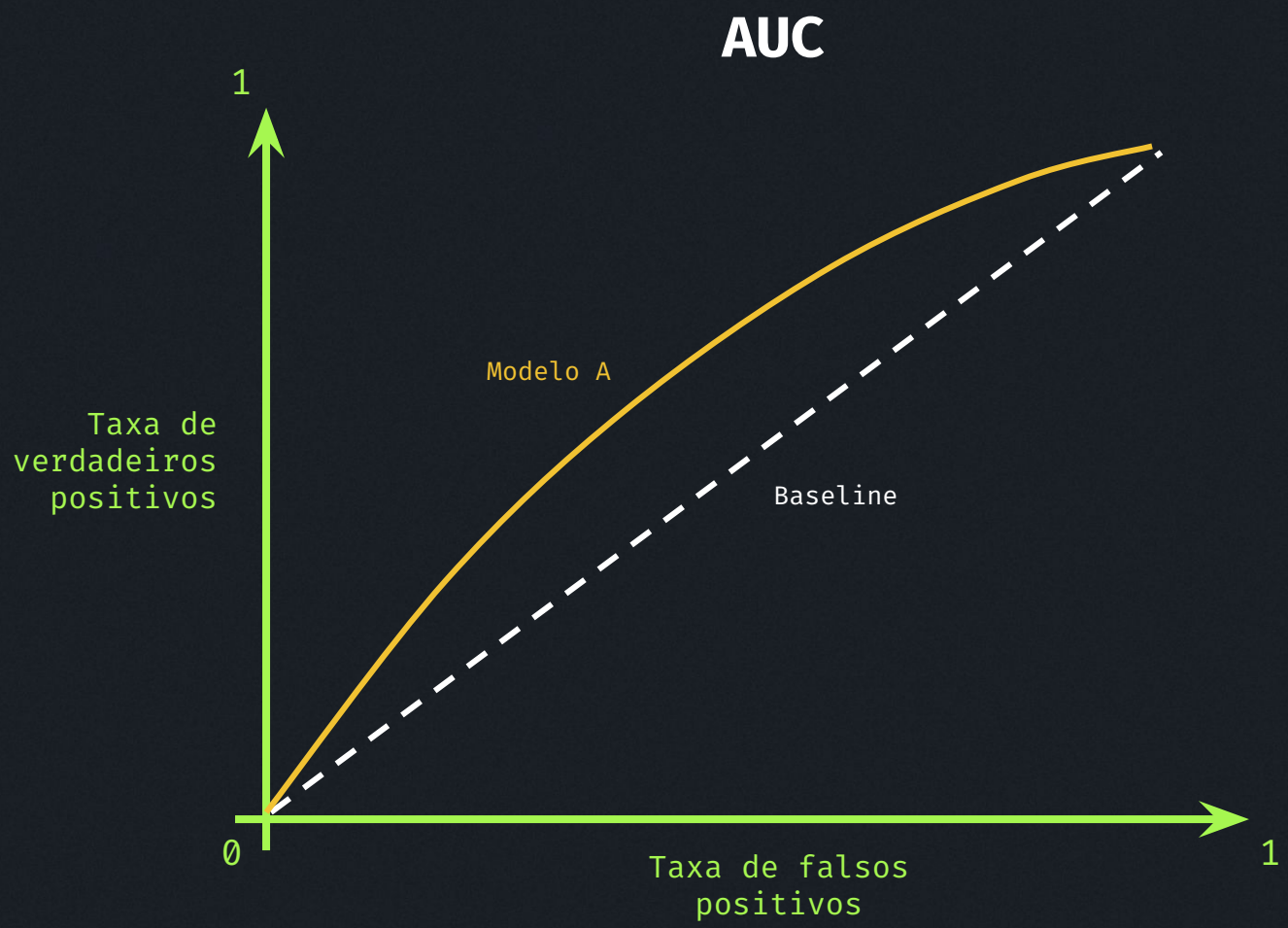
Métrica de avaliação que leva em conta simultaneamente o precision e o recall

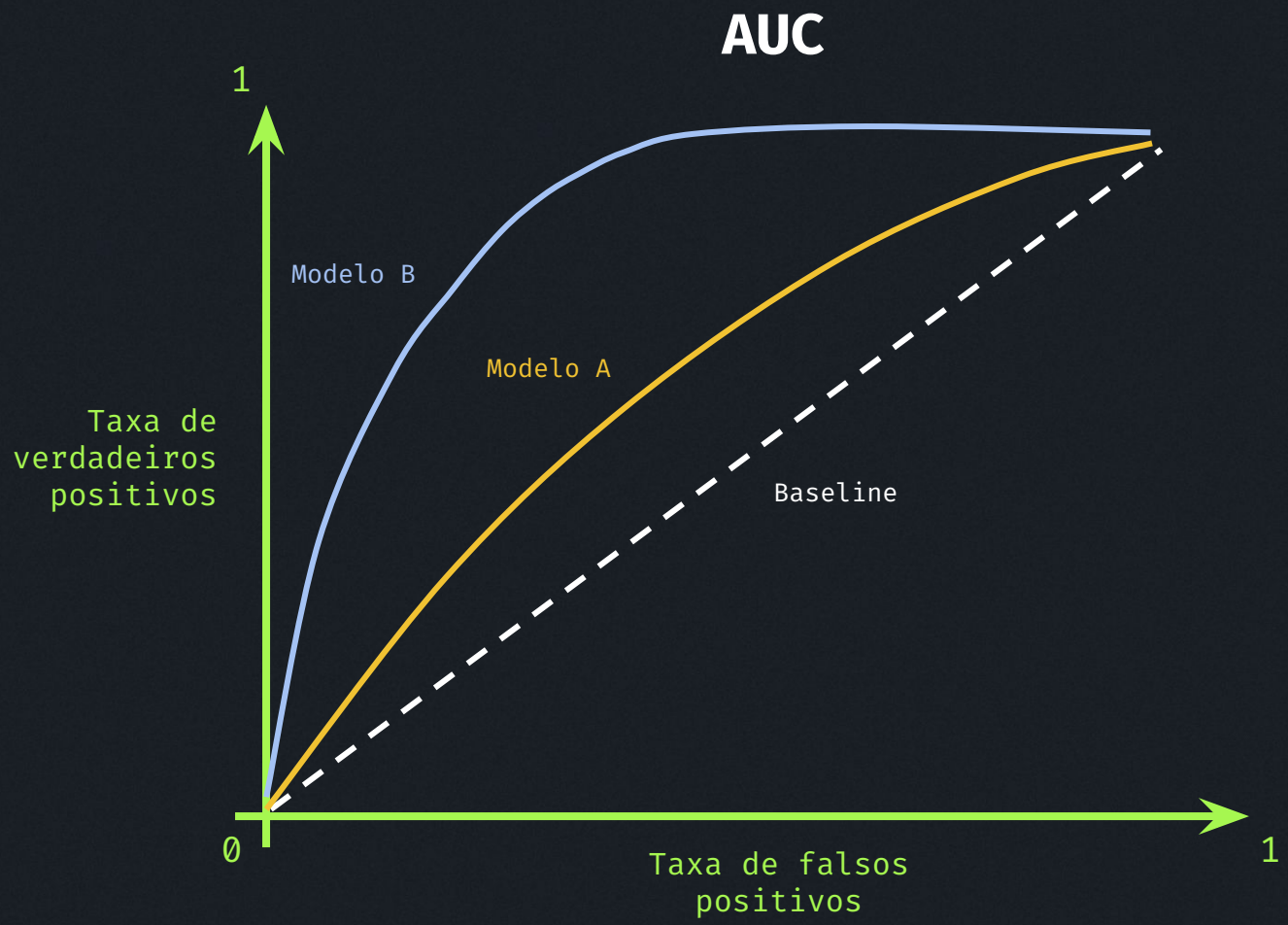
Nota
O F1 score é construído de forma a retornar um valor entre 0 e 1 de modo que este valor seja ponderado pela média aritmética de pior resultado (dentre precision e recall). Assim, não há a possibilidade de uma destas métricas, tendo registrado um valor especialmente excelente, compensar a outra, se esta retornar um valor de avaliação demasiadamente baixo. Valores elevados para o F1 score são obtidos apenas quando precision e recall retornam simultaneamente valores elevados.

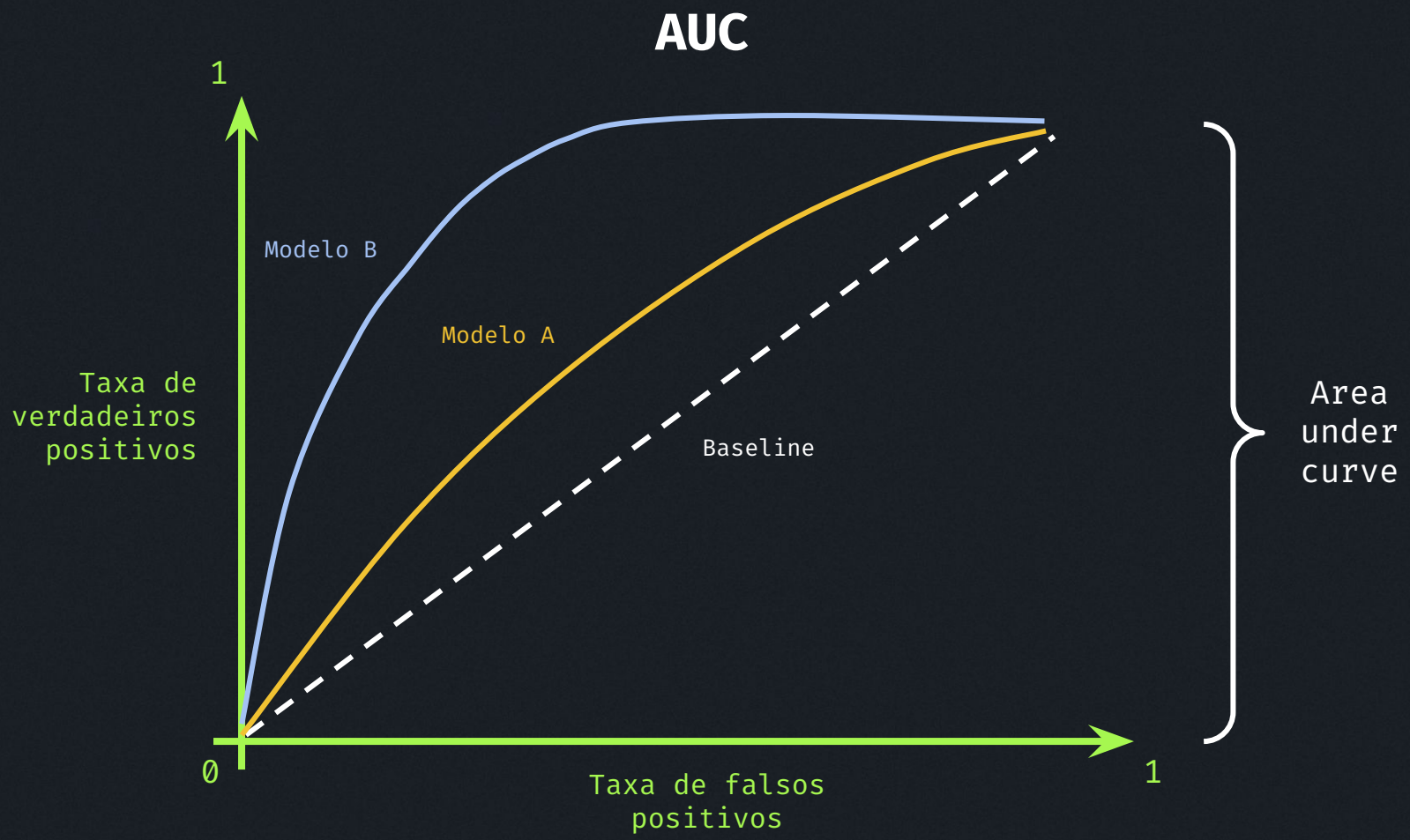
AUC


1
Taxa de verdadeiros positivos
0
1
Taxa de falsos positivos

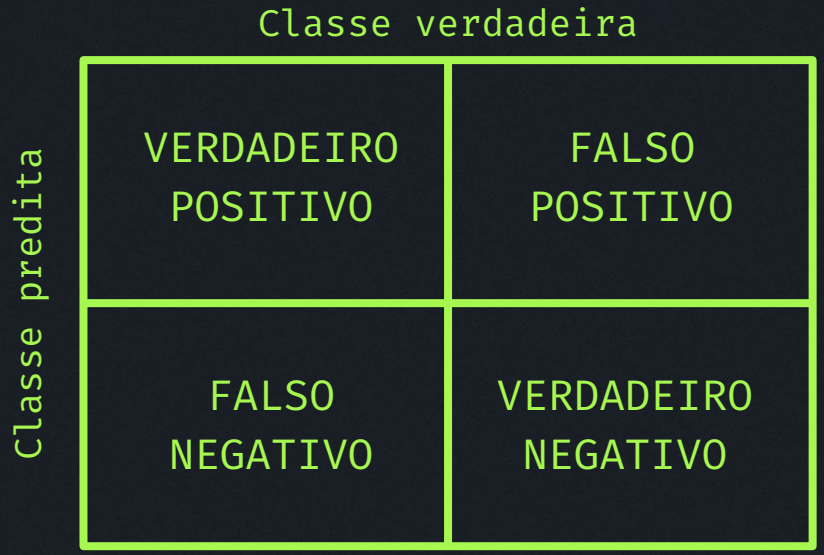

AUC
1
Taxa de verdadeiros positivos
Baseline
0
1
Taxa de falsos positivos


AUC
1
Taxa de verdadeiros positivos
Modelo A
Baseline
0
1
Taxa de falsos positivos


AUC
1
Modelo B
Modelo A
Taxa de verdadeiros positivos
Baseline
0
1
Taxa de falsos positivos


AUC
1
Modelo B
Modelo A
Taxa de verdadeiros positivos
Baseline
Area under curve
0
1
Taxa de falsos positivos

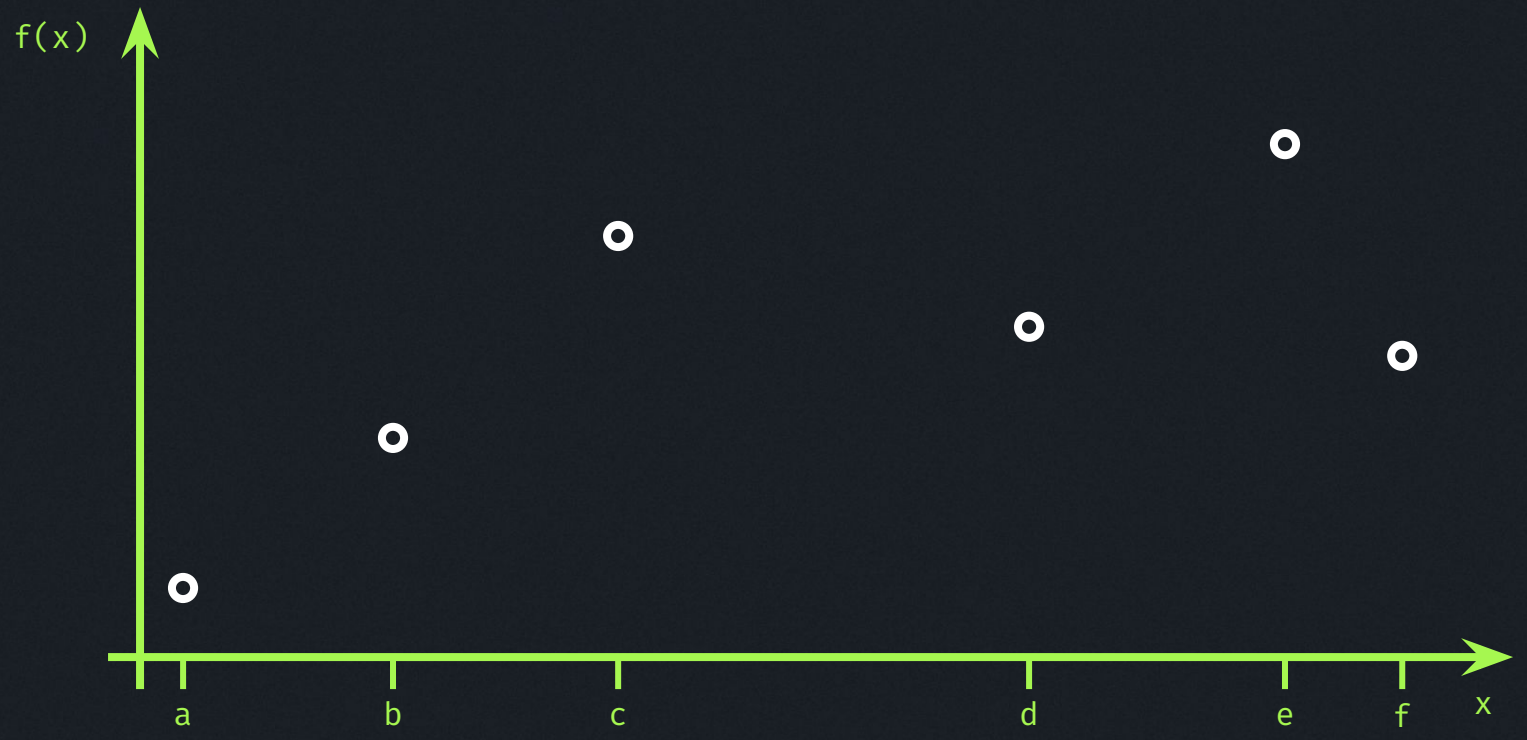
Confusion Matrix



Aprendizado supervisionado: Regressão linear e polinomial

Aprendizado supervisionado: regressão

Definição



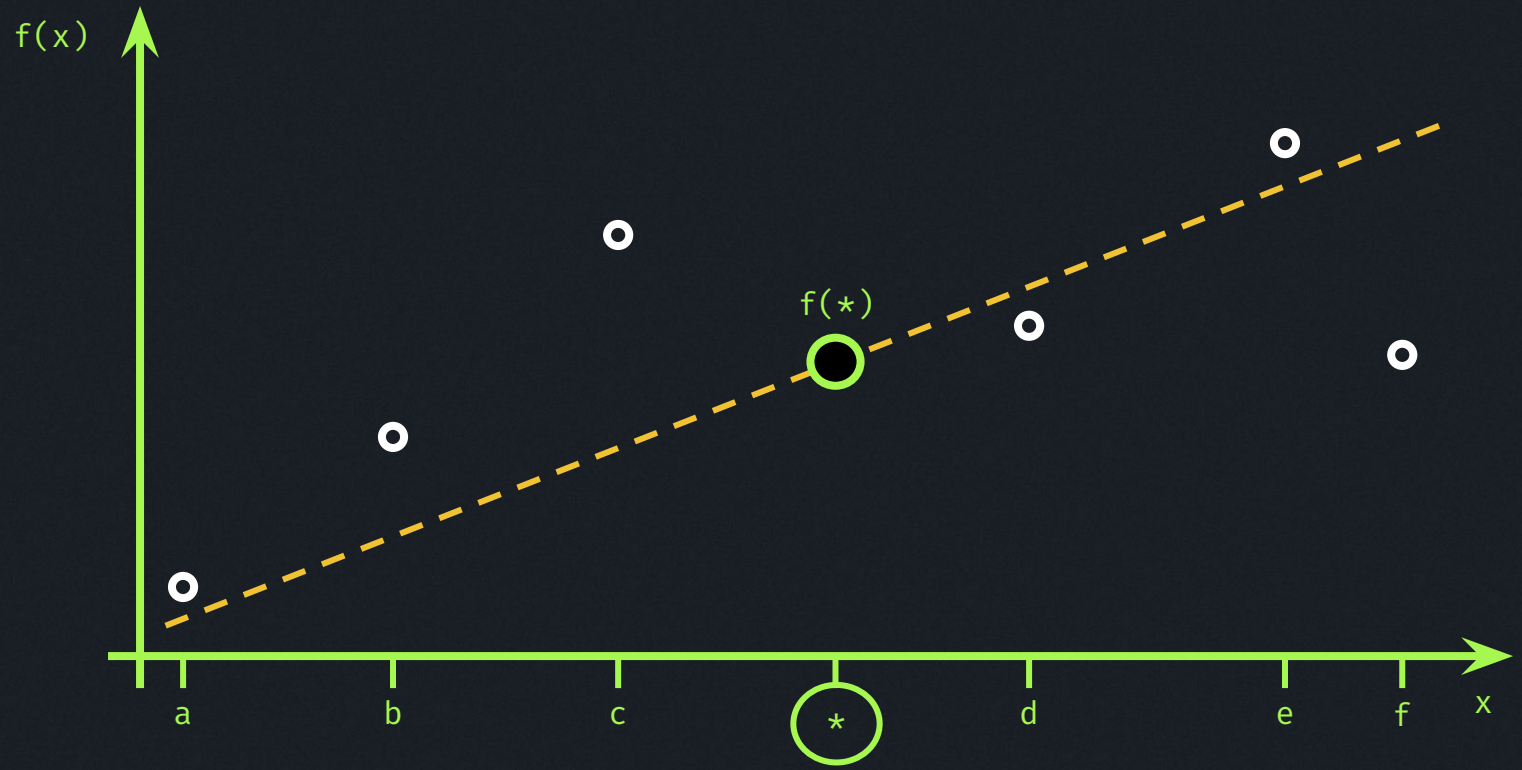
Quando x = *, qual será f(x)?

É possível encontrar uma função que tenha **fitting** com as nossas observações?

Aprendizado supervisionado: regressão

Definição

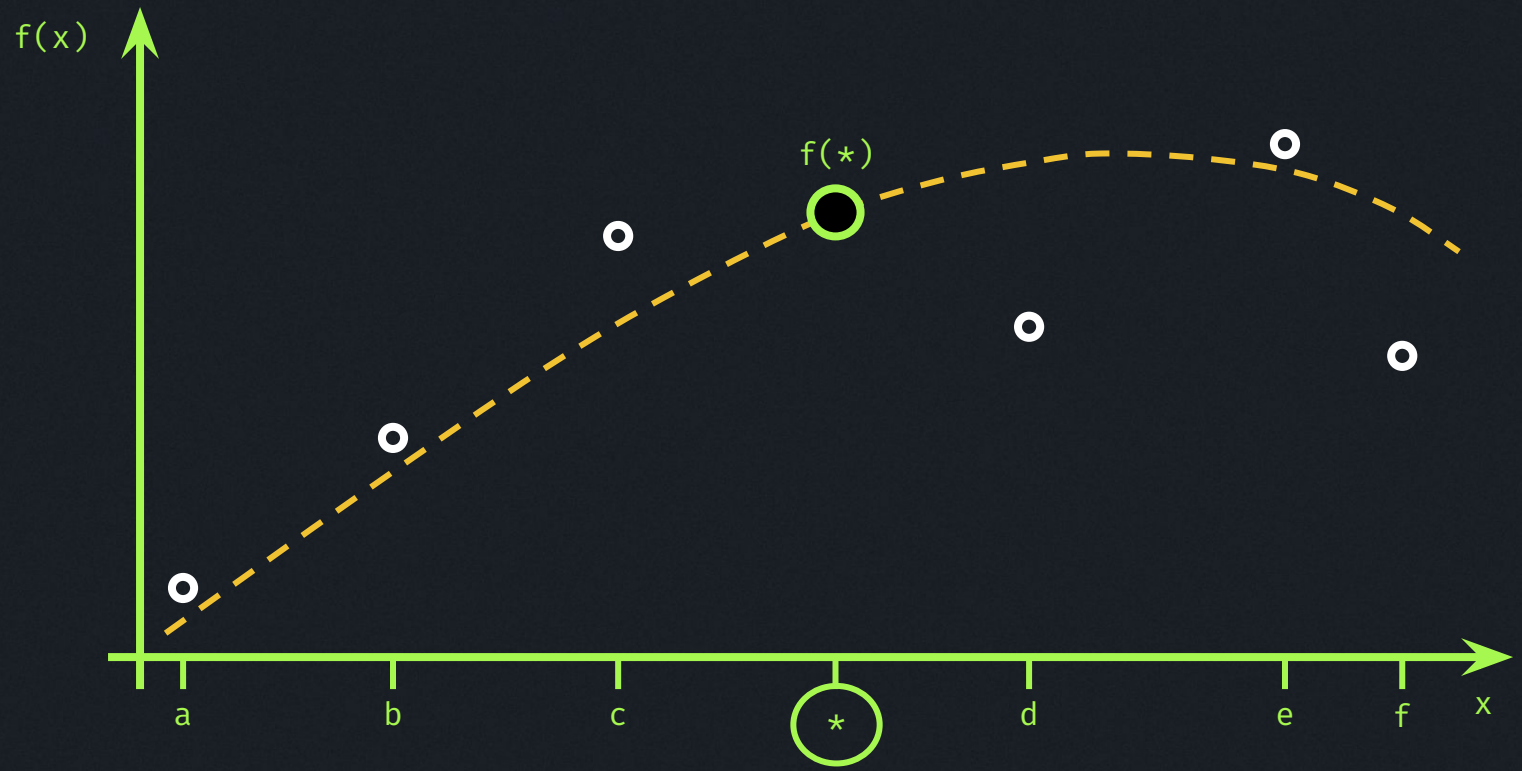
Se fazemos f(x) = ax + b



Aprendizado supervisionado: regressão

Definição

Se fazemos $f(x) = ax^2 + bx + c$

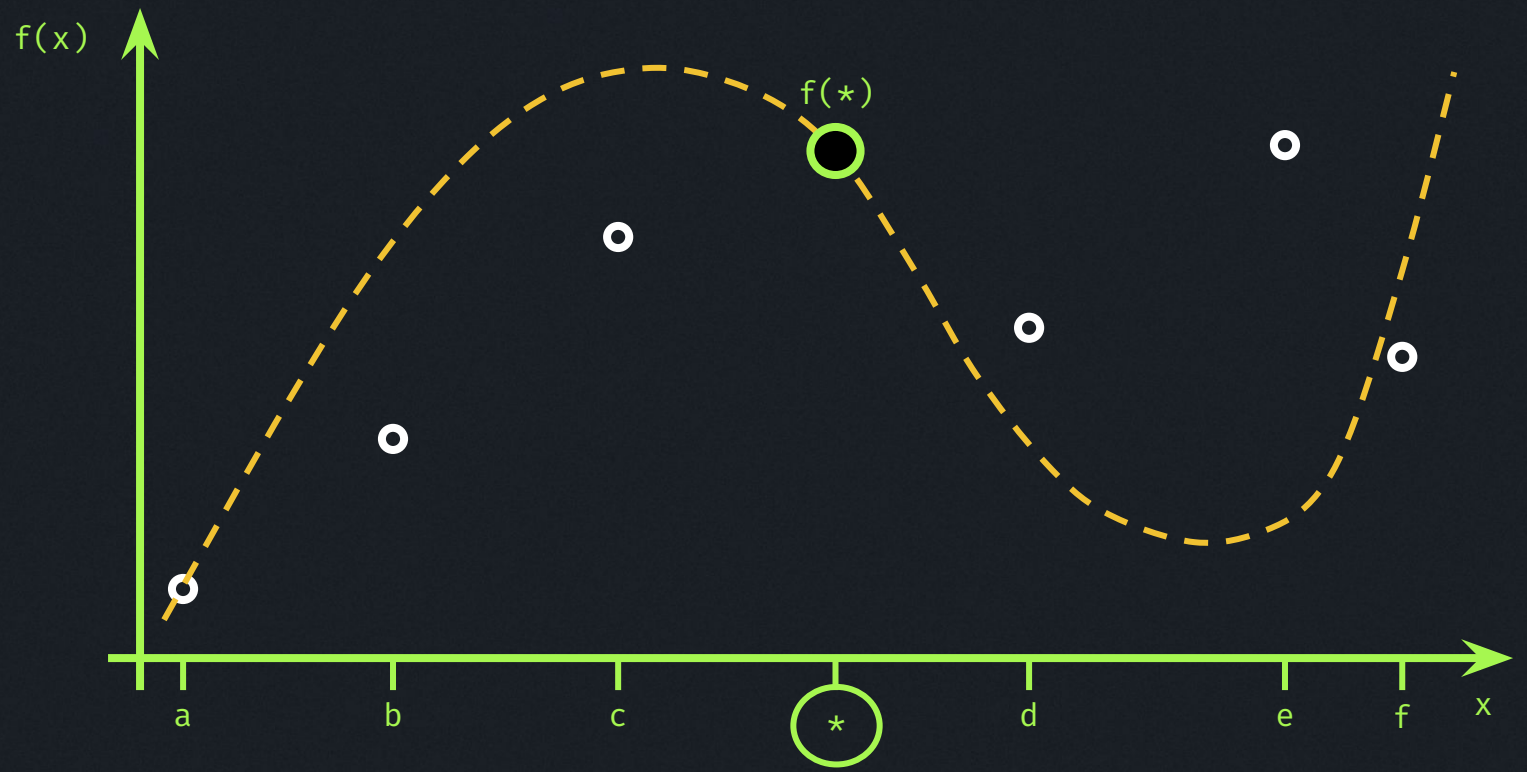


ada

Aprendizado supervisionado: regressão

Definição

Se fazemos $f(x) = ax^3 + bx^2 + cx + d$

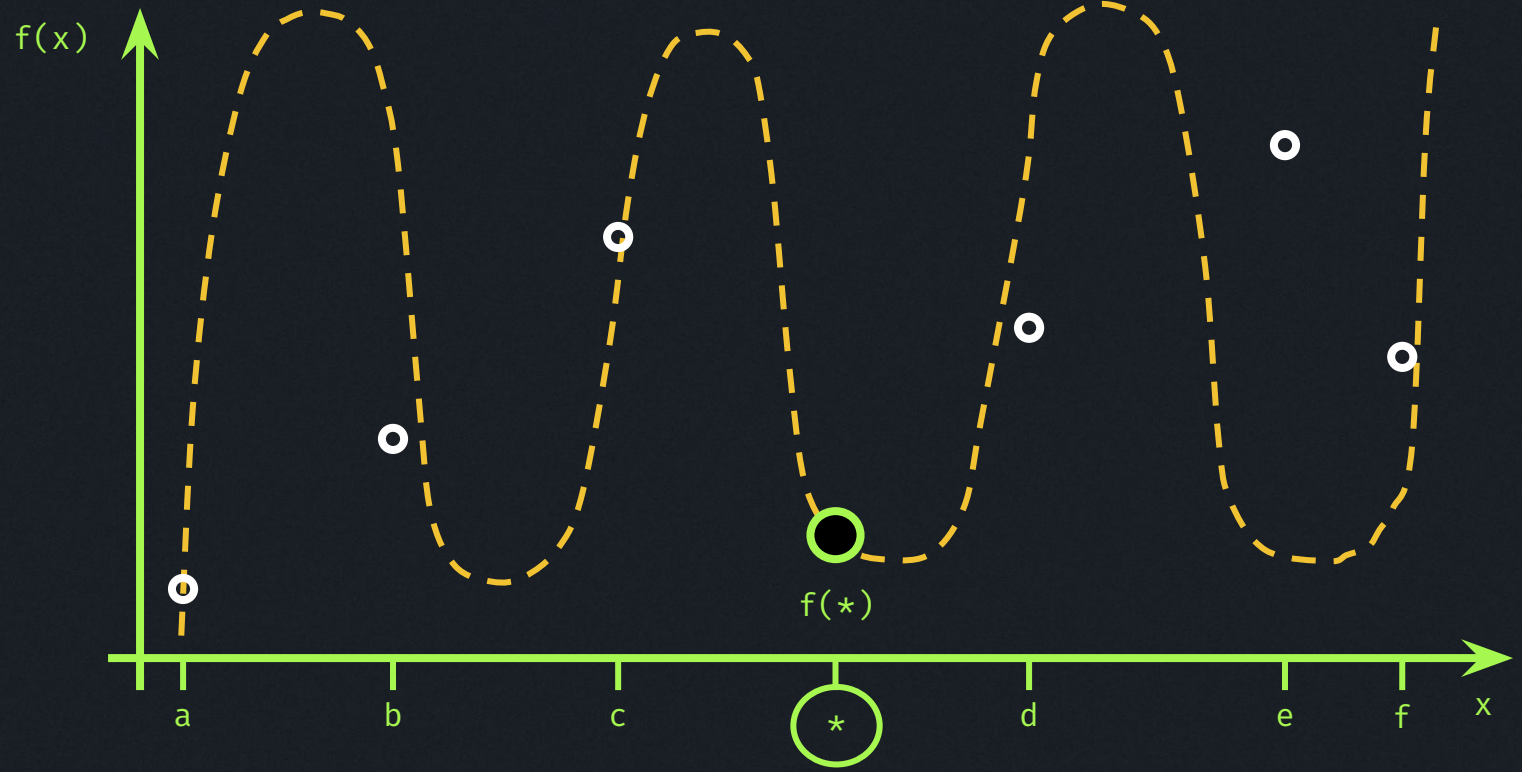


ada

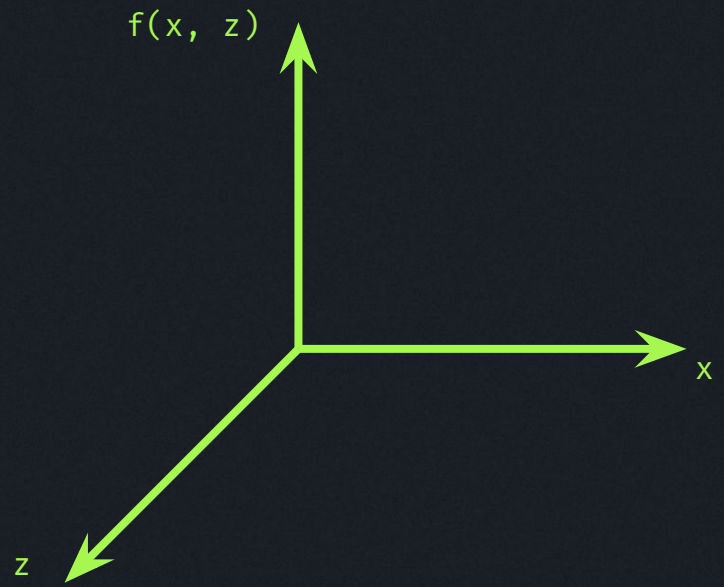
Aprendizado supervisionado: regressão

Definição

Se fazemos $f(x) = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + f$



E como resolveríamos um problema de 3 dimensões?



CEC'14 functions benchmark

desmos.com

Na prática, re-escreveríamos a nossa f(x) = y como f(x, z) = y. Ou então, poderíamos dizer que o x da nossa função f(x) = y seria um vetor composto por dois valores, representado por $\vec{x}$.

Nossa variável resposta dependeria, portanto, de duas outras variáveis, e não de apenas uma. Estaríamos tentando encontrar a função que gera não mais uma linha, mas um **plano** que descreve o nosso conjunto de observações.

Esta abordagem é generalizável para quaisquer n dimensões. Quanto maior n, mais difícil pode ser o problema.

Aprendizado supervisionado: regressão

Definição

Regressão linear, quadrática, cúbica, ...

O grau do polinômio usado para regressão é o que define se a regressão será linear, quadrática, cúbica e assim por diante. Se fazemos:

- $f(x) = ax + b$, temos uma regressão linear
- $f(x) = ax^2 + bx + c$, temos uma regressão quadrática
- $f(x) = ax^3 + bx^2 + cx + d$, temos uma regressão cúbica

Às abordagens quadráticas em diante, podemos nos referir genericamente como **regressões polinomiais**.

Regressão simples e regressão múltipla

Quando $x \in \mathbb{R}$, temos uma regressão simples. Quando $x \in \mathbb{R}^n$, temos uma regressão múltipla. Em outras palavras, quando a variável resposta é dependente de mais de uma variável, deixamos de ter uma regressão simples e passamos a ter uma regressão múltipla. Em termos matemáticos:

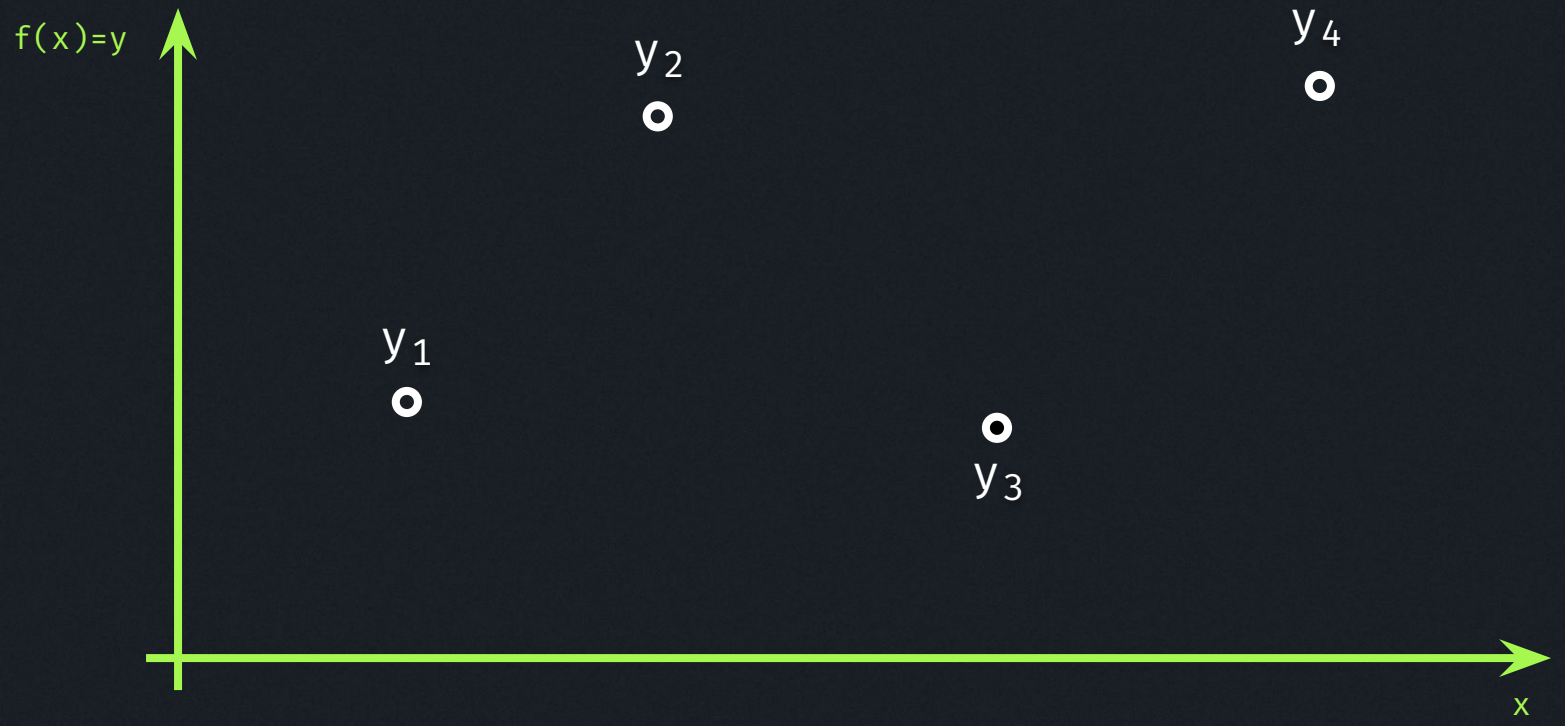
- Se $f(x) = y \mid x \in \mathbb{R}$, temos uma regressão simples
- Se $f(\vec{x}) = y \mid \vec{x} \in \mathbb{R}^n$, temos uma regressão múltipla

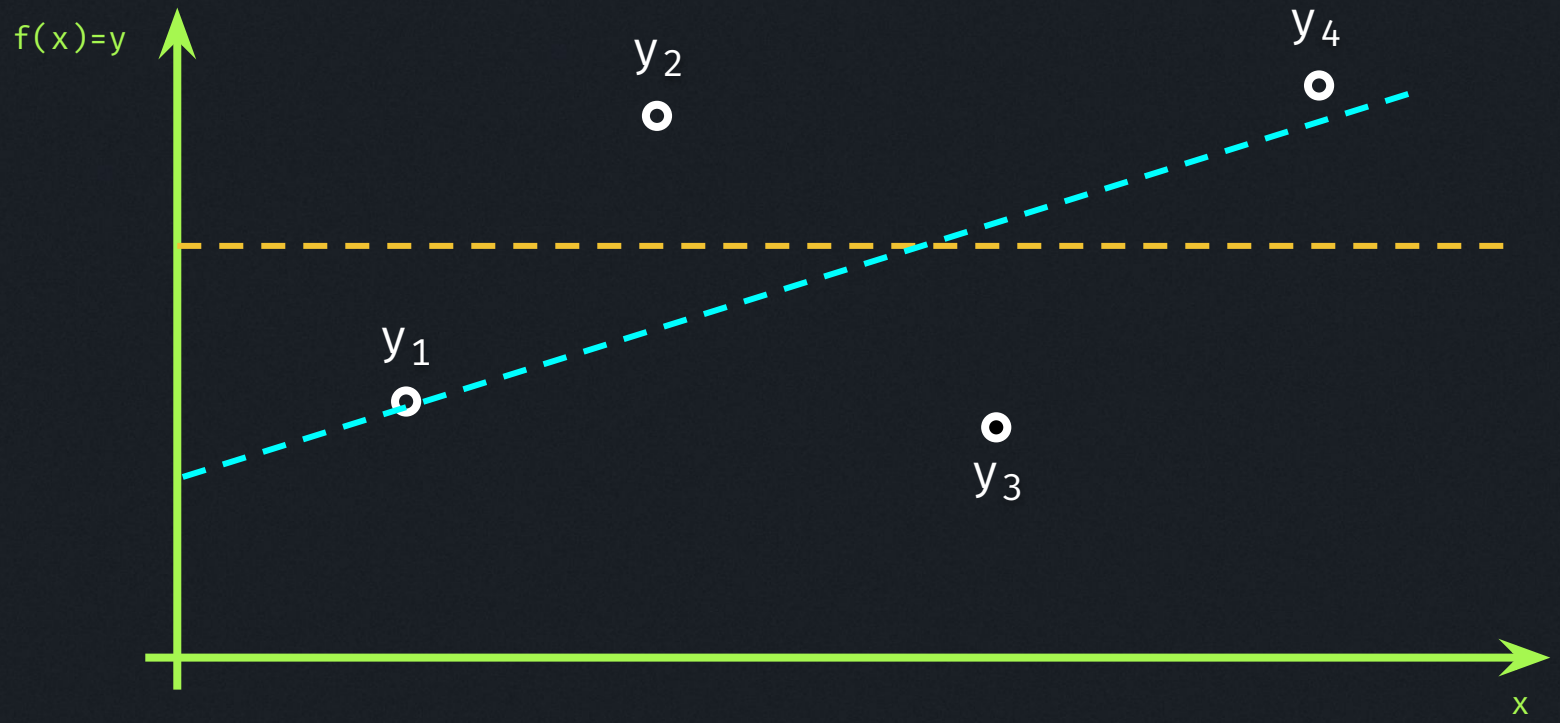
- $f(x) = ax_1 + bx_2 + c$, temos uma reg. lin. múltipla
- $f(x) = ax_1^2 + bx_2^2 + cx_1 + dx_2 + e$

ada

Para avaliarmos um modelo de regressão, precisamos primeiramente definir o que é um erro:

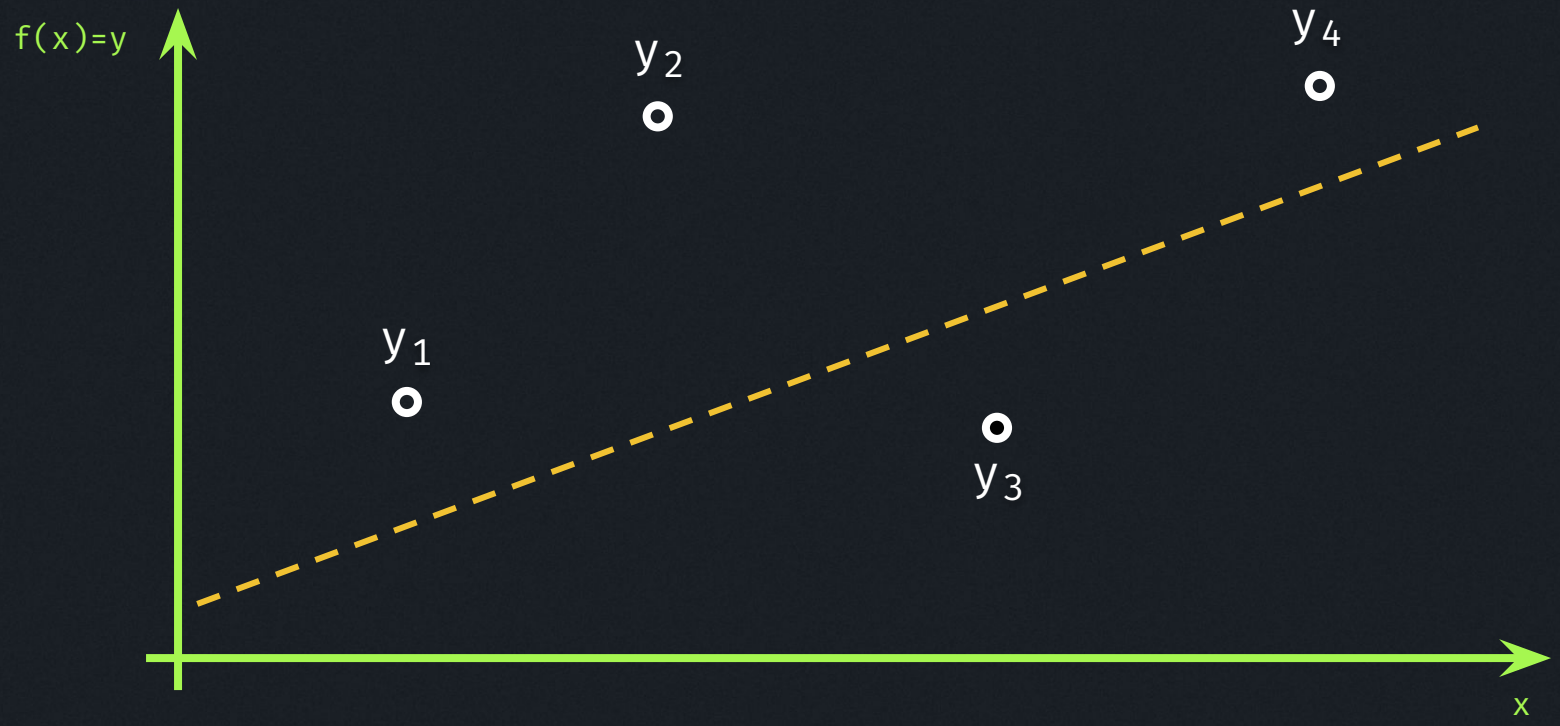
$$e_i = |\hat{y}_i - y_i|$$




f(x)=y
y2
y4
y1
y3
x

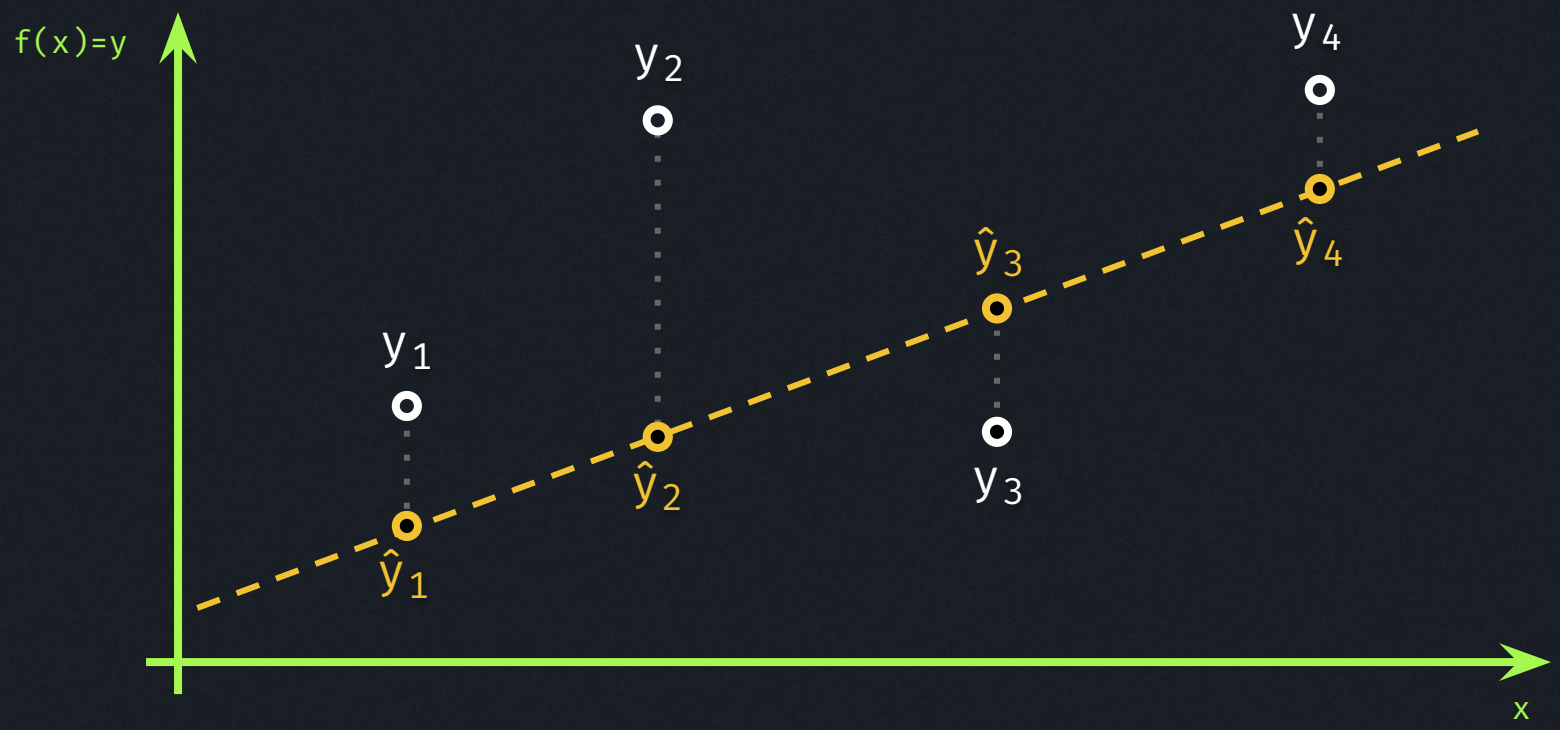
Para avaliarmos um modelo de regressão, precisamos primeiramente definir o que é um erro:

$$e_i = |\hat{y}_i - y_i|$$



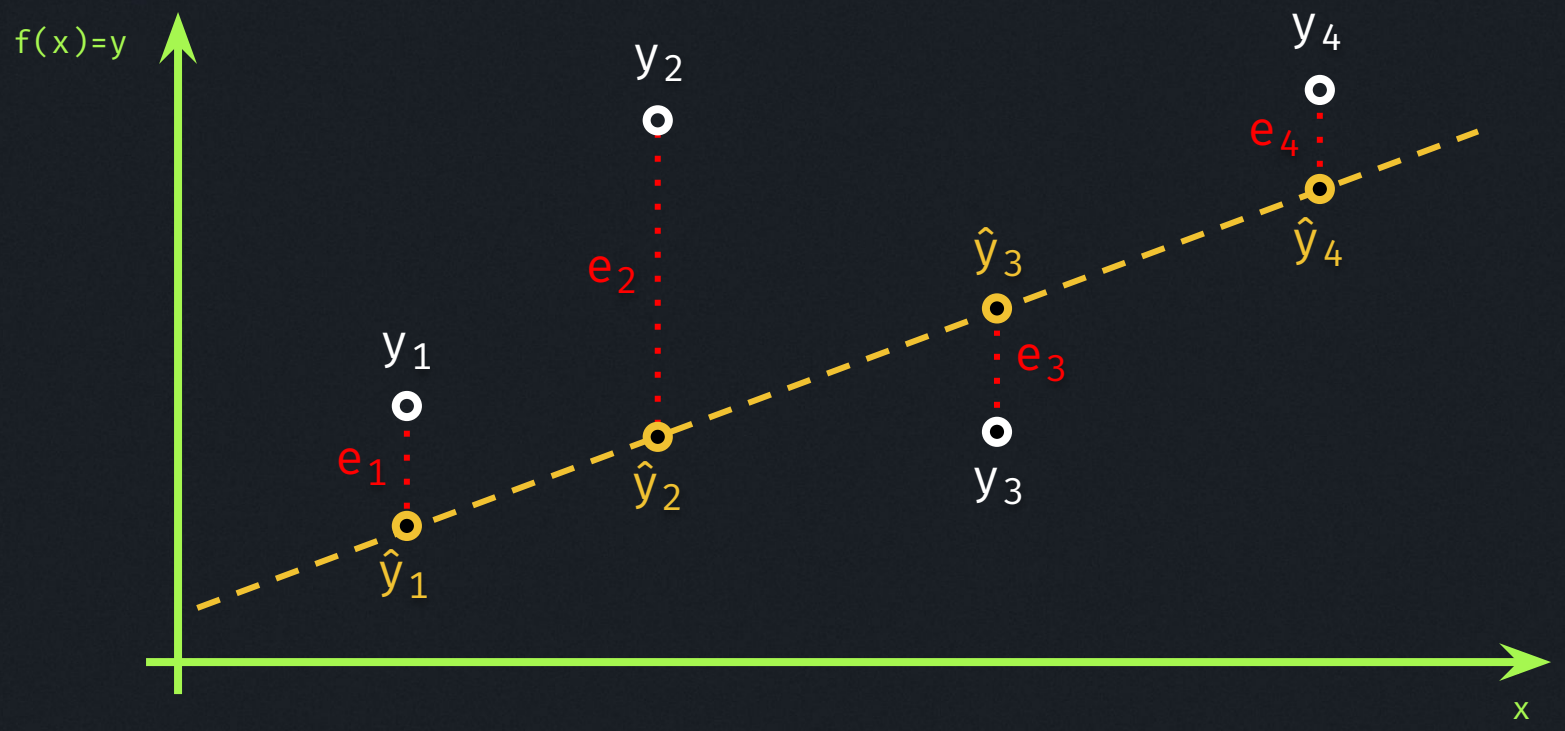
Para avaliarmos um modelo de regressão, precisamos primeiramente definir o que é um erro:

$$e_i = |\hat{y}_i - y_i|$$



Para avaliarmos um modelo de regressão, precisamos primeiramente definir o que é um erro:

$$e_i = |\hat{y}_i - y_i|$$



Mean Absolute Error (MAE)

$$\frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

É a média aritmética dos erros do modelo em valores absolutos.

Útil para obter uma medida de erro com a mesma unidade de medida da variável dependente do problema.

Mas note que errar o preço de uma commodity por 1 dólar será muito mais grave se o preço real da commodity for 2 dólares do que se o preço real for 2000 dólares.

Mean Absolute Percentage Error (MAPE)

$$\frac{\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|}{n} = \frac{\sum_{i=1}^{n} \left| \frac{e_i}{y_i} \right|}{n}$$

É a média aritmética dos erros do modelo em valores percentuais.

Útil para obter uma medida de erro percentual, que indique a relevância da magnitude do erro. Especialmente interessantes também, em casos nos quais podemos ter respostas de magnitudes muito diferentes para diferentes instâncias de entrada.

Mean Squared Error (MSE)

$$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} = \frac{\sum_{i=1}^{n}(e_i)^2}{n}$$

É a média aritmética dos erros do modelo elevados ao quadrado em valores absolutos.

A ideia de elevar os erros ao quadrado é que a penalização por erros maiores que outros se dê exponencialmente. Com o MSE, podemos tolerar com mais facilidade um modelo que comete muitos erros marginais do que um que comete poucos erros muito grandes, por exemplo.

Observe, como ilustração, que $5 - 2 = 3$, enquanto $(5 - 2)^2 = 9$.

Root Mean Squared Error (RMSE)

$$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(e_i)^2}{n}}$$

É a média aritmética dos erros do modelo elevados ao quadrado em valores absolutos.

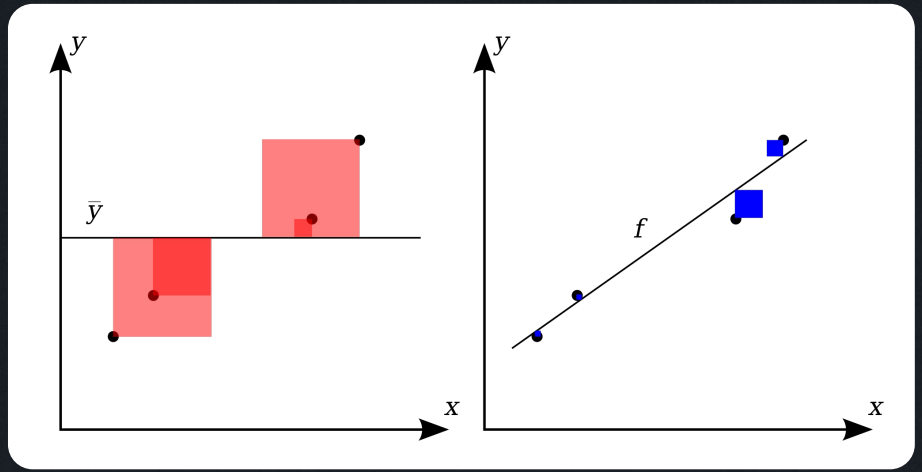
A ideia de elevar os erros ao quadrado é que a penalização por erros maiores que outros se dê exponencialmente. Com o MSE, podemos tolerar com mais facilidade um modelo que comete muitos erros marginais do que um que comete poucos erros muito grandes, por exemplo.

R-squared (R^2)

residual sum of squares

$$1 - \frac{RSS}{TSS}$$

total sum of squares



Mostra o quanto da variância da variável dependente pode ser predita a partir da variável independente. Quanto melhor a regressão linear (direita) se ajusta aos dados em comparação com a média simples (esquerda), mais próximo o valor de R^2 estará de 1.

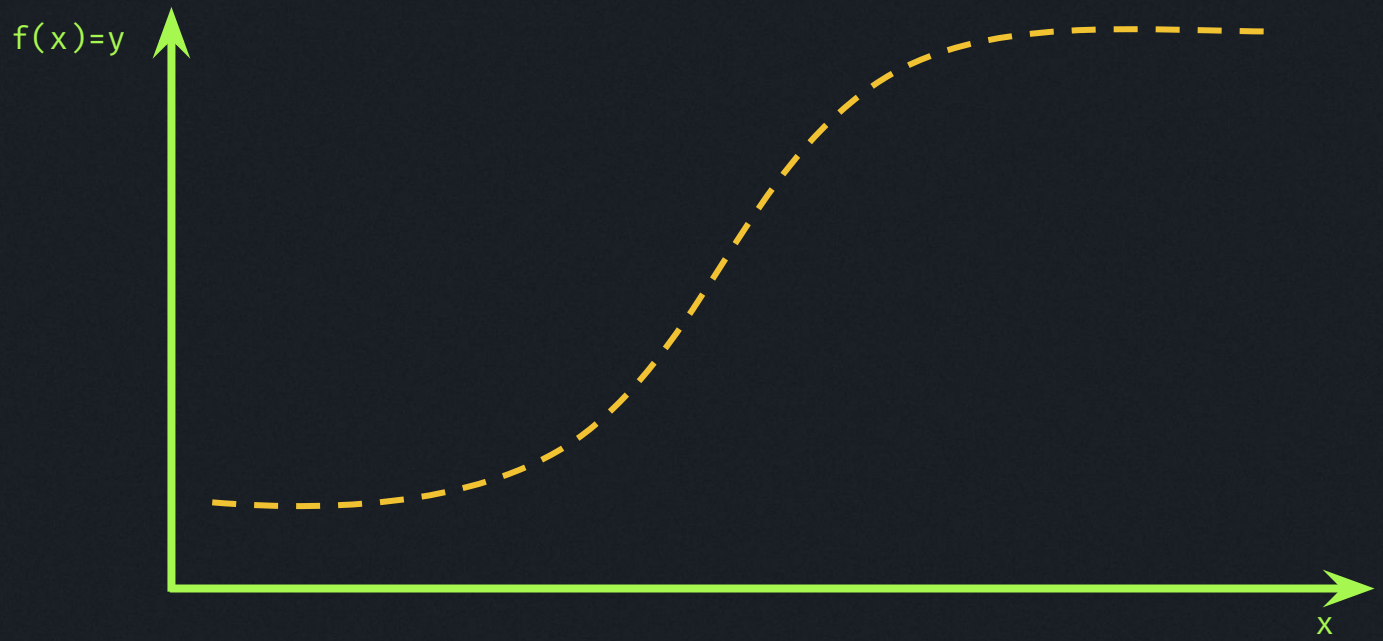
Aprendizado supervisionado: Regressão logística

Função logística

Técnica de classificação baseada na função logística

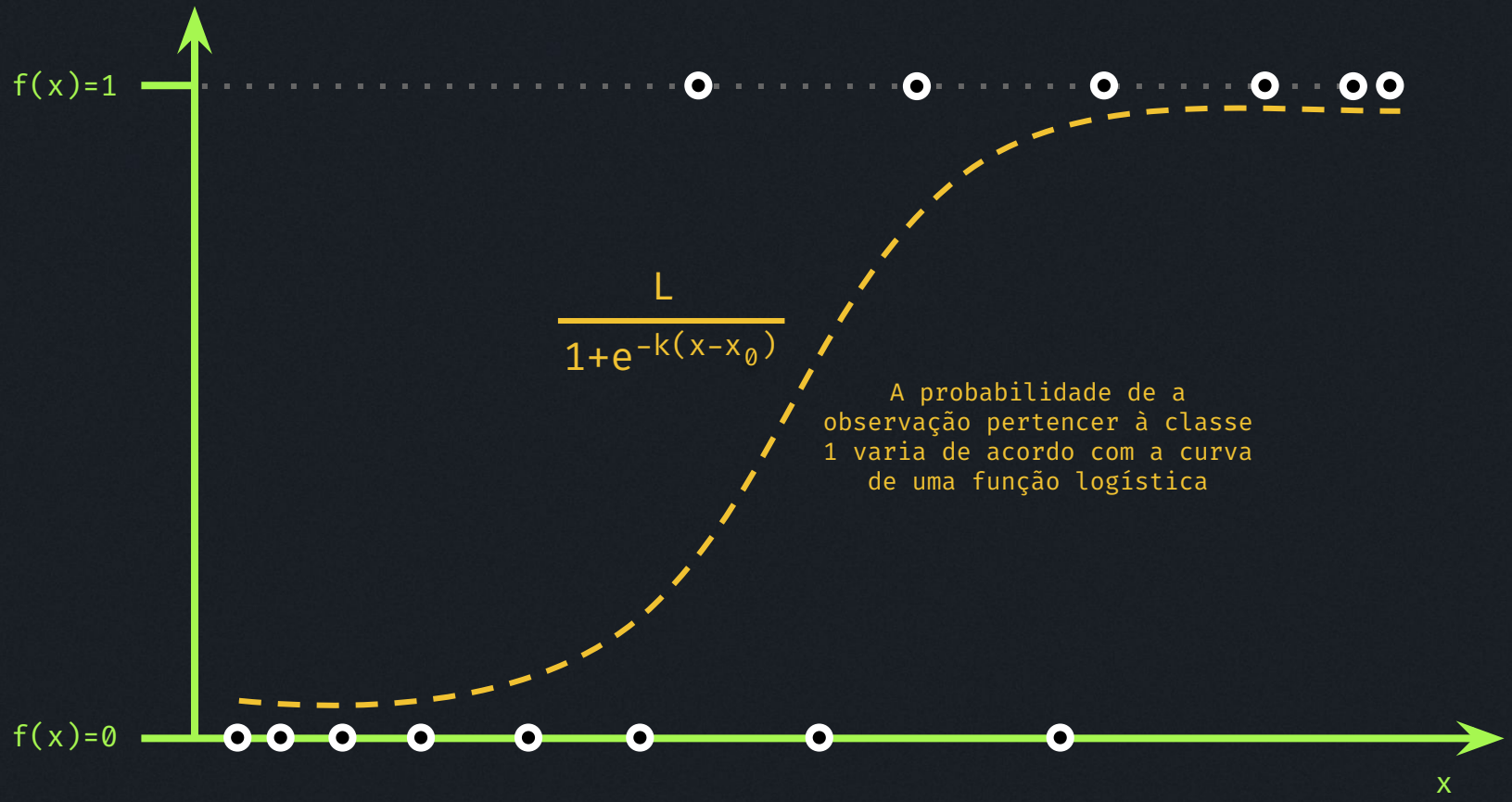
$$f(x) = \frac{L}{1+e^{-k(x-x_0)}}$$

x0: valor de x no ponto médio da curva sigmóide
L: valor máximo da curva
k: declividade da curva



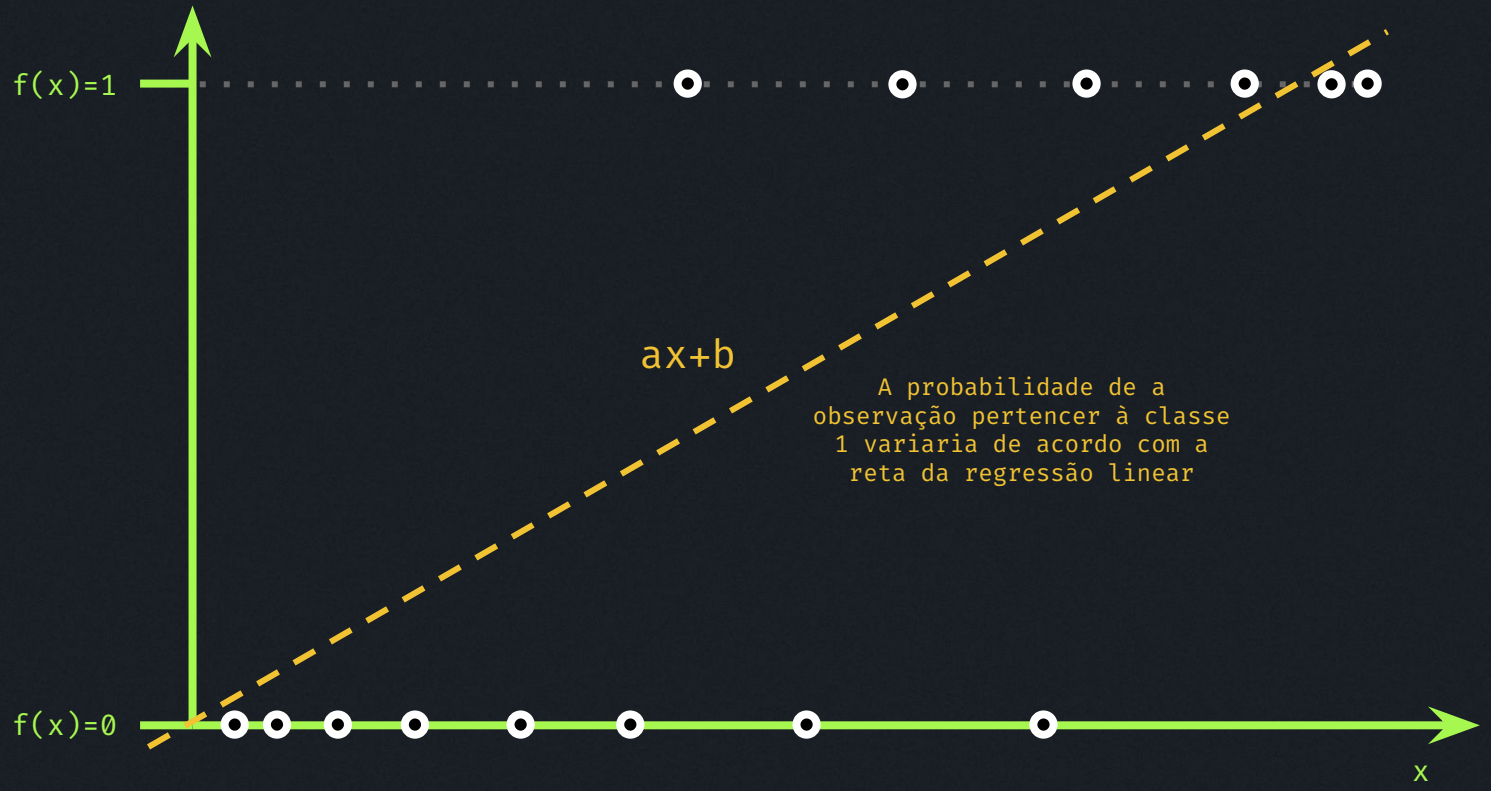
Aprendizado supervisionado: regressão logística

Função logística



ada

Poderíamos usar a equação da reta no lugar da função logística? Se sim, quais seriam as desvantagens?



Regularização

Ordem do polinômio → complexidade do modelo

Complexidade do modelo → overfitting

Em geral:
se eu tenho polinômios de graus diferentes, o polinômio de maior grau gera uma curva mais complexa
Se eu tenho polinômio de mesmo grau, o polinômio com maiores índices, gera uma curva mais complexa

Least Absolute Shrinkage and Selection Operator

$$RSS_{lasso} = \underbrace{\sum_{i=1}^{n} [y_i - (ax_i + b)]^2}_{\text{Função de custo (MSE)}} + \underbrace{\alpha \sum_{j=1}^{p} |w_j|}_{\text{Regularização L1}}$$

Se o nosso objetivo é sempre **MINIMIZAR** a função de custo, então o que significa na prática o termo de regularização, que é uma soma de módulos (sempre positivo, portanto) sendo **SOMADO** ao valor da função de custo

?

Regularização

L1: Lasso regression

Least Absolute Shrinkage and Selection Operator

$$RSS_{lasso} = \underbrace{\sum_{i=1}^{n} [y_i - (ax_i + b)]^2}_{\text{Função de custo (MSE)}} + \underbrace{\alpha \sum_{j=1}^{p} |w_j|}_{\text{Regularização L1}}$$

w é o valor de cada um dos p coeficientes do regressor. Em ax+b, temos a e b, por exemplo.

α (ou λ) é o coeficiente de penalização (ou multiplicador de Lagrange)

Quanto maiores os coeficientes, mais penalizado (desvalorizado) será o resultado obtido pelo modelo. Como o tamanho dos coeficientes é o que dá inclinação e curvaturas ao regressor, a regularização é uma forma de desvalorizar regressores complexos e ousados, para valorizar regressores mais simples e comportados que, em tese, devem ter **maior capacidade de generalização**.

Ridge regression

$$RSS_{lasso} = \underbrace{\sum_{i=1}^{n} [y_i - (ax_i + b)]^2}_{\text{Função de custo (MSE)}} + \underbrace{\alpha \sum_{j=1}^{p} w_j^2}_{\text{Regularização L2}}$$

Qual é o efeito de elevarmos ao quadrado o valor de cada coeficiente do regressor ?

Least Absolute Shrinkage and Selection Operator

$$RSS_{lasso} = \underbrace{\sum_{i=1}^{n} [y_i - (ax_i + b)]^2}_{\text{Função de custo (MSE)}} + \underbrace{\alpha \sum_{j=1}^{p} w_j^2}_{\text{Regularização L2}}$$

Suponha um problema de regressão linear múltipla para o qual obtivemos o regressor
$f(x) = a_1x_1 + a_2x_2 + b$

Suponha para o regressor,
$a_1 = 0.5$
$a_2 = 0.2$
$b = 0.0$

Para a penalidade l1, teríamos:
$0.1 * (|0.5| + |0.2|) = 0.1 * 0.7 = 0.070$

Para a penalidade l2, teríamos:
$0.1 * (0.5^2 + 0.2^2) = 0.1 * 0.29 = 0.029$

A regularização L1 penaliza mais bruscamente e pode apagar a influência de algumas features

Análise estatística de dados numéricos e categóricos

Análise estatística de dados numéricos e categóricos

Testes paramétricos vs não-paramétricos

Testes paramétricos

- Em geral, a **aplicabilidade destes testes é condicionada a suposições sobre a distribuição dos dados**. É comum que seu uso seja restrito a dados que seguem aproximadamente a distribuição normal.
- Fazem uso de **parâmetros populacionais** acerca dos dados, fazendo inferências a partir de métricas como média, mediana, variância, desvio-padrão, correlação, dentre outros.
- Quando as suposições são atendidas, estes testes tem **maior significância estatística.**

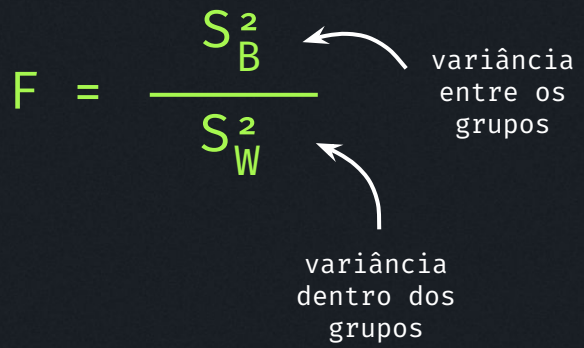
Testes não-paramétricos

- São menos restritivos que os testes paramétricos, com a**plicabilidade independente de condições sobre a distribuição dos dados.** Por isso, inclusive, podem ser aplicados a amostras de dados de tipos mais variados.
- Geralmente são **baseados em ordenações e ranqueamentos**, comparando distribuições e medinas, por exemplo, e não tanto valores brutos.
- **Maior robustez**, já que produzem resultados válidos para uma gama maior de problemas.

ada

Teste não-paramétrico ANOVA - definição

Analysis of Variance



- Usado para aferir se existe diferença estatisticamente significante entre três ou mais grupos de valores independentes.
- Não precisamos assumir uma distribuição específica dos dados.
- Teste de Kruskal-Wallis: comparação das medianas dos grupos em vez das médias para dados ordinais ou que violam as condições do teste paramétrico
- F: evidência contra a hipótese nula de que as amostras vêm da mesma população
- p-valor: probabilidade de obter uma estatística de teste tão extrema quanto a observada caso a hipótese nula seja verdadeira

Feature selection baseada no teste não-paramétrico ANOVA

sklearn.feature_selection
f_classif

- Teste paramétrico baseado no teste ANOVA, para feature selection em problemas de classificação
- Assume distribuição normal dos dados
- Valores maiores de F indicam uma maior relação entre as características e a variável de destino

Teste não-paramétrico qui-quadrado

Chi-squared

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Usado para avaliar a dependência entre duas variáveis categóricas.
- Variáveis não estão relacionadas a uma distribuição específica e não são numéricas.
- Usado para responder perguntas como:
  - Existe uma relação entre o gênero e a preferência de um determinado produto?
  - Existe uma relação entre o tipo de tratamento médico e a recuperação de pacientes?
  - Existe uma associação entre a idade e a opinião política?
- Quanto maior o valor qui-quadrado observado, maior a divergência entre as variações observadas e as variações esperadas sob a hipótese nula de independência das variáveis e, portanto, maior a evidência contrária a esta hipótese.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

Feature Selection vs Dimensionality Reduction

Principal Component Analysis (PCA)
Formulação e uso do método

- **Pré-processamento dos dados:**
  Normalizar as características para média zero e variância unitária.

- **Cálculo da matriz de covariância:**
  Calcular a matriz de covariância dos dados.

- **Autovalores e autovetores:**
  Calcular os autovalores e autovetores da matriz de covariância.

- **Ordenação dos componentes principais:**
  Ordenar os componentes principais em ordem decrescente com base nos autovalores.

- **Seleção dos componentes principais:**
  Selecionar os primeiros k componentes principais com os maiores autovalores.

- **Projeção dos dados:**
  Projetar os dados originais no espaço dos componentes principais selecionados.

Principal Component Analysis (PCA)
Avaliação de componentes do PCA

- **Contribuição para a variabilidade:**
  Componentes principais com maiores autovalores têm maior contribuição para a variabilidade dos dados.

- **Importância das características:**
  Os autovetores indicam a importância relativa das características originais para cada componente principal.

- **Relações entre características:**
  Os sinais dos autovetores revelam relações positivas ou negativas entre as características originais.

- **Visualização:**
  Projetar os dados no espaço dos componentes principais permite identificar padrões e agrupamentos.