Aprendizagem de Máquina

Introdução e Conceitos Fundamentais

Telmo de Menezes e Silva Filho
tmfilho@gmail.com/telmo@de.ufpb.br
www.de.ufpb.br

UFPB

Departamento de ESTATÍSTICA

Sumário

Sobre a Disciplina

- Aulas às segundas e quartas
  - 8h-10h
  - Link da aula no Google Meet (ou YouTube) estará disponível no SIGAA no dia anterior
  - A aula será sempre gravada e disponibilizada posteriormente
  - Alunos podem fazer perguntas ao longo da aula (ficarei logado em dois dispositivos para acompanhar isso)

Sobre a Disciplina

- Avaliação:
  - Uma prova (05/10) – Peso 1
  - Um projeto em grupo – Peso 2
    - Grupos e temas serão definidos após a prova
    - Projetos poderão envolver soluções para desafios do kaggle, benchmarks para novos conjuntos, etc
    - Criatividadeh
    - Projeto será entregue em repositório público no Github
  - O projeto e as questões da prova que peçam código poderão ser desenvolvidos em qualquer linguagem de programação

Sobre a Disciplina

- Como praticar?
  - Após cada aula, aproveitem que o vídeo ficará online para revisar
  - Quaisquer notebooks associados às aulas também ficarão disponíveis
  - Após cada conteúdo, escolham problemas conhecidos e simples e apliquem o que foi visto em sala
    - Usem esses miniprojetos para tirar dúvidas, deixando-os disponíveis no Github

Sobre a Disciplina

- Referências (disponíveis na biblioteca virtual):
  - FACELI, Katti; LORENA, Ana Carolina; GAMA, João. Inteligência Artificial:Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: LTC, 2015; 2019. 378p. ISBN: 9788521618805.
  - HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. The elements of statistical learning: data mining, inference, and prediction. 2.ed. Nova Yorque: Springer, c2009. 745 p. (Springer series in statistics) ISBN: 9780387848570.

Conceitos Fundamentais

Probabilidade Condicional

Probabilidade Condicional

Considere o seguinte exemplo em que temos 100 dias, em que pode ter **chovido ou não** em uma cidade (C e C^c, respectivamente) e que podem ter tido **umidade relativa do ar média alta (acima de 80%) ou não** (U e U^c, respectivamente)

	U	U^c	
C	55	5	60
C^c	20	20	40
	75	25	100

Probabilidade Condicional

Qual é a probabilidade de Chover?

	U	U^c	
C	**55**	**5**	**60**
C^c	20	20	40
	75	25	**100**

$$P(C) = \frac{60}{100}$$

Probabilidade Condicional

Qual é a probabilidade de que a umidade relativa média do ar em um dia seja alta?

	U	U^c	
C	**55**	5	60
C^c	**20**	20	40
	75	25	**100**

$$P(U) = \frac{75}{100}$$

Probabilidade Condicional

Qual é a probabilidade de que chova **E** a umidade relativa média do ar seja alta?

	U	U^c	
C	**55**	5	60
C^c	20	20	40
	75	25	**100**

$$P(U \cap C) = \frac{55}{100}$$

Probabilidade Condicional

Dado que choveu, qual a probabilidade de a umidade tenha sido alta? Em outras palavras, entre os dias em que choveu, qual foi a proporção de dias muito úmidos?

	U	U^c	
C	**55**	5	**60**
C^c	20	20	40
	75	25	100

$$P(U|C) = \frac{P(U \cap C)}{P(C)} = \frac{55}{60}$$

Probabilidade Condicional

Dado que choveu, qual a probabilidade de a umidade tenha sido alta? Em outras palavras, entre os dias em que choveu, qual foi a proporção de dias muito úmidos?

Isso é o que chamamos de **Probabilidade Condicional** e é um conceito importantíssimo para a Aprendizagem de Máquina

	U	U^c	
C	**55**	5	**60**
C^c	20	20	40
	75	25	100

$$P(U|C) = \frac{P(U \cap C)}{P(C)} = \frac{55}{60}$$

Teorema de Bayes

Quando conhecemos a probabilidade condicional de um evento dados outros eventos **mutuamente exclusivos** e sabemos que o primeiro ocorreu, temos como calcular as probabilidades condicionais “inversas”

Em outras palavras, se eu sei $P(U|C)$ e $P(U|C^c)$ e U aconteceu, eu posso calcular $P(C|U)$ e $P(C^c|U)$

Para isso usamos um “truque de mestre”, chamado **Teorema de Bayes**

Teorema de Bayes

Quando conhecemos a probabilidade condicional de um evento dados outros eventos **mutuamente exclusivos** e sabemos que o primeiro ocorreu, temos como calcular as probabilidades condicionais “inversas”

$$P(C|U) = \frac{P(C \cap U)}{P(U)} = \frac{P(C \cap U)}{P(U \cap C) + P(U \cap C^c)} = \frac{55}{55 + 20} = \frac{55}{75}$$

	U	U^c	
C	55	5	60
C^c	20	20	40
	75	25	100

Teorema de Bayes

O Teorema é útil quando não temos as informações completas em uma tabelinha. Por exemplo: suponha que sabemos que $P(C) = 60/100$, $P(U|C) = 55/60$ e $P(U|C^c) = 20/40$.

$$P(C|U) = \frac{P(C \cap U)}{P(U)} = \frac{P(C \cap U)?}{P(U \cap C)? + P(U \cap C^c)?}$$

Da definição de probabilidade condicional:

$$P(C|U) = \frac{P(C \cap U)}{P(U)} = \frac{P(C \cap U)}{P(U \cap C) + P(U \cap C^c)} = \frac{P(U|C)P(C)}{P(U|C)P(C) + P(U|C^c)P(C^c)}$$

Teorema de Bayes

O Teorema é útil quando não temos as informações completas em uma tabelinha. Por exemplo: suponha que sabemos que $P(C) = 60/100$, $P(U|C) = 55/60$ e $P(U|C^c) = 20/40$.

$$P(C|U) = \frac{P(U|C)P(C)}{P(U|C)P(C) + P(U|C^c)P(C^c)} = \frac{\frac{55}{\cancel{60}} \cdot \frac{\cancel{60}}{100}}{\frac{55}{\cancel{60}} \cdot \frac{\cancel{60}}{100} + \frac{20}{\cancel{40}} \cdot \frac{\cancel{40}}{100}}$$

$$P(C|U) = \frac{\frac{55}{\cancel{100}}}{\frac{55}{\cancel{100}} + \frac{20}{\cancel{100}}} = \frac{55}{75}$$

Teorema de Bayes

O Teorema é útil quando não temos as informações completas em uma tabelinha. Por exemplo: suponha que sabemos que $P(C) = 60/100$, $P(U|C) = 55/60$ e $P(U|C^c) = 20/40$.

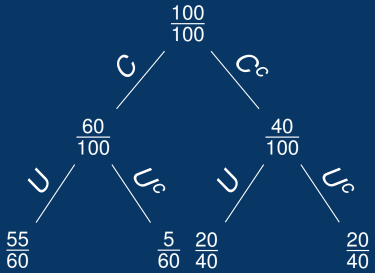
$$P(C|U) = \frac{\color{red}{P(U|C)P(C)}}{\color{red}{P(U|C)P(C)} + P(U|C^c)P(C^c)} = \frac{\frac{55}{\cancel{60}} \cdot \frac{\cancel{60}}{100}}{\frac{55}{\cancel{60}} \cdot \frac{\cancel{60}}{100} + \frac{20}{\cancel{40}} \cdot \frac{\cancel{40}}{100}}$$

$$P(C|U) = \frac{\frac{55}{\cancel{100}}}{\frac{55}{\cancel{100}} + \frac{20}{\cancel{100}}} = \frac{55}{75}$$

Note a **repetição** do termo do numerador no denominador

Teorema de Bayes

Essas informações são comumente visualizadas na forma de uma árvore. Note que com isso, já conseguimos tomar decisões baseadas em evidências.



$$P(C|U) = \frac{\frac{55}{60} \cdot \frac{60}{100}}{\frac{55}{60} \cdot \frac{60}{100} + \frac{20}{40} \cdot \frac{40}{100}}$$

Eventos Independentes

Sejam A e B eventos quaisquer. A e B são ditos **eventos independentes** se, e somente se,

$$P(A \cap B) = P(A) \cdot P(B)$$

Prova: $P(A \cap B) = P(A|B) \cdot P(B) = P(A) \cdot P(B)$

Nota: Parafraseando, temos que se o evento A é independente do evento B, quer dizer que a probabilidade de A não é afetada pela ocorrência ou não-ocorrência do evento B. Isso é uma suposição frequente dos algoritmos de aprendizagem de máquina em relação às variáveis do problema (mais a seguir)

Variáveis Aleatórias

Variáveis Aleatórias

Definição: Seja E um experimento aleatório qualquer e Ω o espaço amostral associado ao experimento E. Uma função X que associe cada elemento $\omega \in \Omega$ um número real $X(\omega)$ é denominada de **variável aleatória**.

Variáveis Aleatórias

Para esse curso, podemos entender variável aleatória como qualquer característica numérica associada a um experimento aleatório.

Exemplo 1: Consideremos o experimento aleatório E que consiste em observar a ocorrência ou não de chuva ao longo de 100 dias. Temos assim que,

$$Y = \{0, 1\}$$

Y é **discreta** e podemos modelar seu comportamento usando uma **função de probabilidade**

Variáveis Aleatórias

Para esse curso, podemos entender variável aleatória como qualquer característica numérica associada a um experimento aleatório.

Exemplo 2: Consideremos o experimento aleatório E que consiste em observar a umidade relativa do ar média ao longo de 100 dias. Temos assim que,

$$X \in [0, 1]$$

X é contínua e podemos modelar seu comportamento usando uma **função densidade de probabilidade - fdp**

Variável Aleatória Discreta

Seja Y uma variável aleatória. Se o número de valores possíveis de Y for finito ou enumerável, diremos que Y é uma **variável aleatória discreta**. Além disso, poderemos atribuir medida de probabilidade aos seus valores possíveis y_i, com $i \geq 1$. A cada possível resultado, teremos uma probabilidade $P(Y = y_i)$, que satisfaz as seguintes propriedades:

i) $P(Y = y_i) \geq 0$;

ii) $\sum_{i=1}^{\infty} P(Y = y_i) = 1$.

Observação: P é chamado de função de probabilidade uma vez que Y é uma variável aleatória discreta.

Variável Aleatória Contínua

Seja X uma variável aleatória. Diremos que X é uma **variável aleatória contínua** se existir uma **função densidade de probabilidade** (**fdp**) de X que satisfaz as seguintes condições:

i) $f(x) \geq 0$, para todo valor de x,

ii) $\int_{-\infty}^{+\infty} f(x)dx = 1$,

iii) para quaisquer a e b tal que $-\infty < a < b < +\infty$, teremos que

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

Note que P(X = x) = 0

Por que eu preciso saber disso?

- Na aprendizagem de máquina, trabalhamos com dados
- Os dados observados não correspondem a todos os possíveis valores dos atributos de interesse
- Podemos também assumir que certos valores ou intervalos de valores observados dos atributos são mais prováveis, dependendo dos padrões que podem ser encontrados nos dados
- Daí podemos usar variáveis aleatórias para modelar os nossos atributos

Distribuições de Probabilidade

Distribuições de Probabilidade

- Normalmente nós desconhecemos as reais funções de probabilidade e de densidade de probabilidade das nossas variáveis
- No entanto, para diversos casos do dia-a-dia mais conhecidos, temos modelos prontos, chamados de **distribuições de probabilidade**
- Esses modelos fornecem funções de probabilidade e fdps conhecidas, além de média (esperança) e variância

Distribuição de Bernoulli

Usada para modelar sucesso/fracasso, sim/não, falso/verdadeiro, choveu/não choveu ou qualquer outra variável aleatória discreta em que haja **dicotomia**.

$$Y = \begin{cases} 0 & \text{se fracasso;} \\ 1 & \text{se sucesso.} \end{cases}$$

Distribuição de Bernoulli

Seja Y uma variável aleatória discreta tal que $Y \sim Bernoulli(p)$. Então, a **função de probabilidade** é dada por:

$$P(Y = y) = p^y(1 - p)^{1-y},$$

com $0 < p < 1$ e $y \in \{0, 1\}$, sendo a probabilidade de sucesso $P(Y = 1) = p$ e seu complementar a probabilidade de fracasso, isto é, $P(Y = 0) = 1 - p$.

Além disso, o valor esperado (média) da variável aleatória Y é p e sua variância é $p(1 - p)$, isto é, $E(Y) = p$ e $Var(Y) = p(1 - p)$, respectivamente.

Distribuição Normal

- A **distribuição normal** também é chamada de **distribuição Gaussiana** e tem a forma de um **sino**.
- A distribuição normal é uma das distribuições mais importantes da Estatística.
- Diversos resultados na inferência estatística se apoiam no uso dessa distribuição.

Distribuição Normal

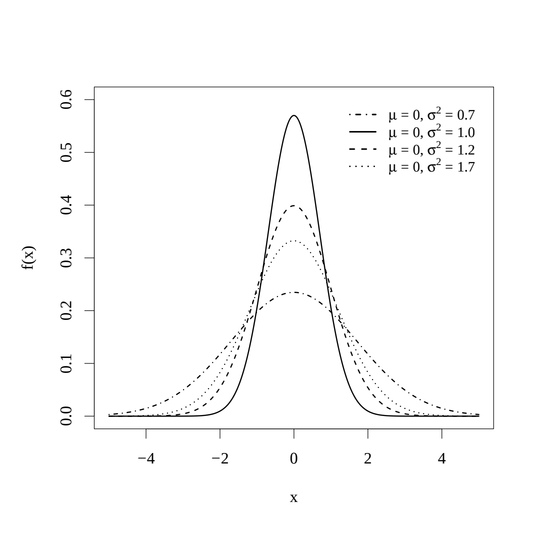
Uma variável aleatória X (contínua) que assume valores em $-\infty < X < \infty$ tem distribuição normal ou (Gaussiana) se sua função densidade de probabilidade (parametrizada por média μ e variância σ^2) é dada por:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

com $x \in \mathbb{R}$, $\mu \in \mathbb{R}$ e $\sigma^2 \geq 0$

Observação: Se X é uma variável aleatória normalmente distribuída de parâmetros μ e σ^2, denotaremos esse fato por $X \sim \mathcal{N}(\mu, \sigma^2)$

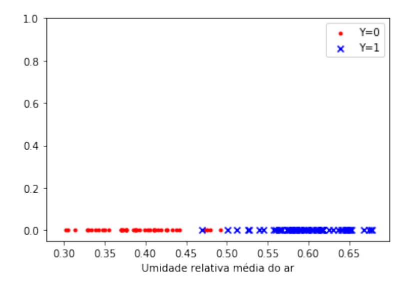
Distribuição Normal



Tudo Junto e Misturado

Como usamos tudo isso?

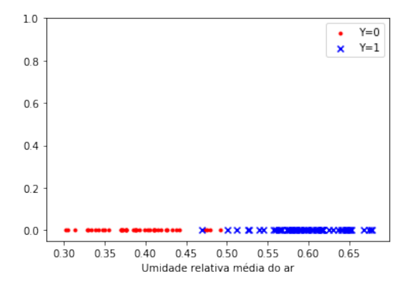
- Vamos voltar pra o nosso problema da chuva
- Imagine que nossos dados observados tomem a seguinte forma:



- Aqui, temos 60% dos dados representando dias de chuva
- Podemos modelar isso como uma variável $Y \sim \mathit{Bernoulli}(0{,}6)$

Como usamos tudo isso?

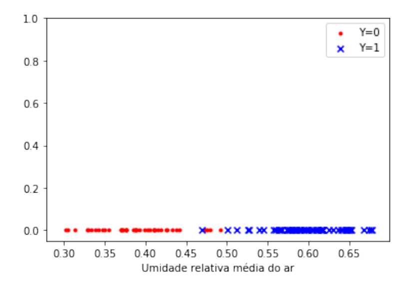
- Vamos voltar pra o nosso problema da chuva
- Imagine que nossos dados observados tomem a seguinte forma:



- O eixo X representa a umidade relativa média do ar em cada um dos dias
  - Essa variável é contínua
  - Vamos modelar esses dados usando duas distribuições normais, condicionadas ao valor de Y

Como usamos tudo isso?

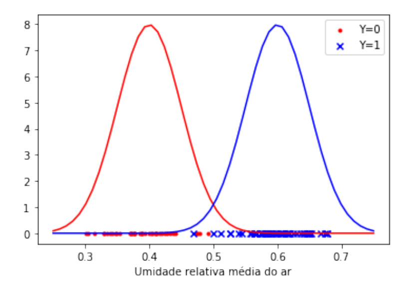
- Vamos voltar pra o nosso problema da chuva
- Imagine que nossos dados observados tomem a seguinte forma:



- O eixo X representa a umidade relativa média do ar em cada um dos dias
  - $X|Y = 0 \sim \mathcal{N}(0{,}4, 0{,}05^2)$
  - $X|Y = 1 \sim \mathcal{N}(0{,}6, 0{,}05^2)$

Como usamos tudo isso?

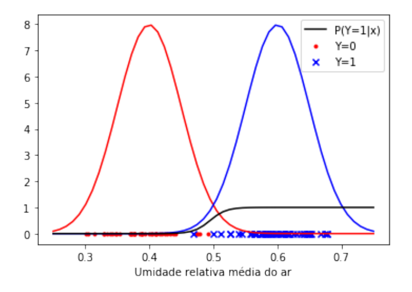
- Vamos voltar pra o nosso problema da chuva
- Imagine que nossos dados observados tomem a seguinte forma:

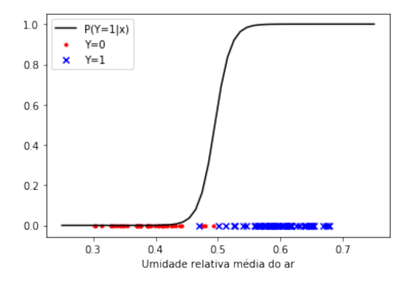


- O eixo X representa a umidade relativa média do ar em cada um dos dias
  - $X|Y = 0 \sim \mathcal{N}(0{,}4, 0{,}05^2)$
  - $X|Y = 1 \sim \mathcal{N}(0{,}6, 0{,}05^2)$

Como usamos tudo isso?

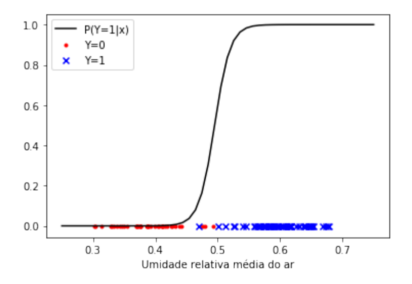
- Dado um novo dia com umidade relativa do ar x, qual é a chance de chover $P(Y = 1|x)$?





Como usamos tudo isso?

- Como calculamos $P(Y = 1|x)$?



- Usamos o **Teorema de Bayes**

$$P(Y = 1|x) = \frac{f_{X|Y=1}(x)P(Y = 1)}{f_{X|Y=1}(x)P(Y = 1) + f_{X|Y=0}(x)P(Y = 0)}$$

- Em que

$$f_{X|Y=1}(x) = \frac{1}{\sqrt{2\pi}0{,}05} \exp\left\{-\frac{(x - 0{,}6)^2}{2 \cdot 0{,}05^2}\right\}$$

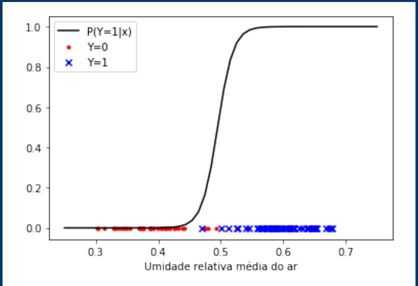
$$f_{X|Y=0}(x) = \frac{1}{\sqrt{2\pi}0{,}05} \exp\left\{-\frac{(x - 0{,}4)^2}{2 \cdot 0{,}05^2}\right\}$$

$$P(Y = 1) = 0{,}6$$

$$P(Y = 0) = 0{,}4$$

Como usamos tudo isso?

- Como calculamos $P(Y = 1|x)$?



- Usamos o **Teorema de Bayes**

$$P(Y = 1|x) = \frac{f_{X|Y=1}(x) \cdot 0{,}6}{f_{X|Y=1}(x) \cdot 0{,}6 + f_{X|Y=0}(x) \cdot 0{,}4}$$

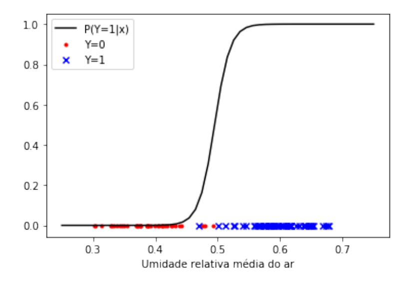
- Em que

$$f_{X|Y=1}(x) = \frac{1}{\sqrt{2\pi}0{,}05} \exp\left\{-\frac{(x-0{,}6)^2}{2 \cdot 0{,}05^2}\right\}$$

$$f_{X|Y=0}(x) = \frac{1}{\sqrt{2\pi}0{,}05} \exp\left\{-\frac{(x-0{,}4)^2}{2 \cdot 0{,}05^2}\right\}$$

Como usamos tudo isso?

- Como calculamos $P(Y = 1|x)$?

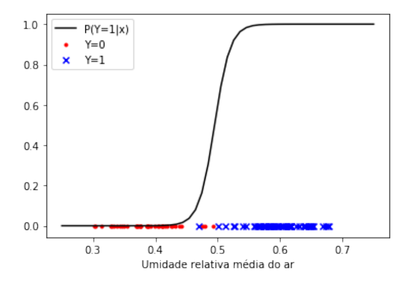


- Usamos o **Teorema de Bayes**

$$P(Y=1|x) = \frac{\cancel{\frac{1}{\sqrt{2\pi}0,05}} \exp\left\{-\frac{(x-0,6)^2}{2\cdot 0,05^2}\right\}\cdot 0,6}{\cancel{\frac{1}{\sqrt{2\pi}0,05}} \exp\left\{-\frac{(x-0,6)^2}{2\cdot 0,05^2}\right\}\cdot 0,6 + \cancel{\frac{1}{\sqrt{2\pi}0,05}} \exp\left\{-\frac{(x-0,4)^2}{2\cdot 0,05^2}\right\}\cdot 0,4}$$

Como usamos tudo isso?

- Como calculamos $P(Y = 1|x)$?



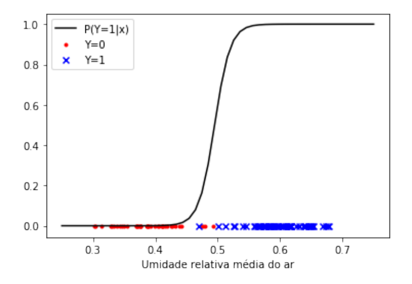
- Usamos o **Teorema de Bayes**

$$P(Y = 1|x) = \frac{\exp\left\{-\frac{(x-0{,}6)^2}{2\cdot 0{,}05^2}\right\}\cdot 0{,}6}{\exp\left\{-\frac{(x-0{,}6)^2}{2\cdot 0{,}05^2}\right\}\cdot 0{,}6 + \exp\left\{-\frac{(x-0{,}4)^2}{2\cdot 0{,}05^2}\right\}\cdot 0{,}4}$$

- Dividindo tudo pelo numerador:

Como usamos tudo isso?

- Como calculamos $P(Y = 1|x)$?

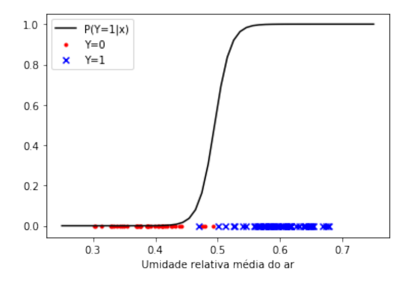


- Usamos o **Teorema de Bayes**

$$P(Y = 1|x) = \frac{1}{1 + \dfrac{\exp\left\{-\dfrac{(x-0,4)^2}{2\cdot 0,05^2}\right\}\cdot 0,4}{\exp\left\{-\dfrac{(x-0,6)^2}{2\cdot 0,05^2}\right\}\cdot 0,6}}$$

Como usamos tudo isso?

- Como calculamos $P(Y = 1|x)$?

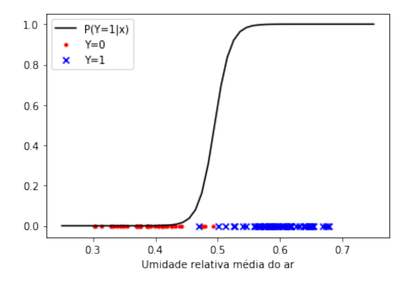


- Usamos o **Teorema de Bayes**

$$P(Y = 1|x) = \frac{1}{1 + \exp\left\{\frac{(x-0{,}6)^2-(x-0{,}4)^2}{0{,}005}\right\} \cdot \frac{2}{3}}$$

Como usamos tudo isso?

- Como calculamos $P(Y = 1|x)$?



- Usamos o **Teorema de Bayes**

$$P(Y = 1|x) = \frac{1}{1 + \exp\left\{\frac{x^2 - 1{,}2x + 0{,}36 - x^2 + 0{,}8x - 0{,}16}{0{,}005}\right\} \cdot \frac{2}{3}}$$

$$P(Y = 1|x) = \frac{1}{1 + \exp\left\{\frac{-0{,}4x + 0{,}2}{0{,}005}\right\} \cdot \frac{2}{3}}$$

$$P(Y = 1|x) = \frac{1}{1 + \frac{2}{3} \cdot e^{-(80x - 40)}}$$

Para terminar

- Parabéns, você acabou de entender o seu primeiro modelo de Aprendizagem de Máquina!
- Essa é a base do famoso **Naïve Bayes**
- Nesse caso, como sabemos exatamente quais as distribuições dos dados, temos um modelo **Bayes ótimo**
- O formato final do cálculo da probabilidade lembrou outro modelo bastante conhecido. Você sabe qual é?

Para terminar

- Como veremos mais à frente no curso, acabamos de resolver um problema de **classificação**
- Chamamos os diferentes valores de Y de classes
- A cada valor da função de probabilidade de Y, i.e. $P(Y=1)$ e $P(Y=0)$ damos o nome de **priori de classe** (*class prior*)
  - Costumamos estimá-los usando os dados observados
- As probabilidades $P(Y=1|x)$ são chamadas de probabilidades a **posteriori de classe**
- Os valores de média e de desvio/variância das Normais condicionadas a cada classe são estimados usando os dados observados de cada classe

Aprendizagem de Máquina

Introdução e Conceitos Fundamentais

Telmo de Menezes e Silva Filho
tmfilho@gmail.com/telmo@de.ufpb.br
www.de.ufpb.br

UFPB

Departamento de ESTATÍSTICA