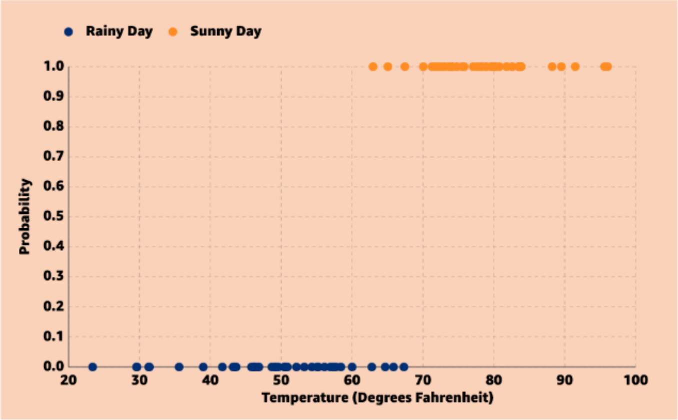
# Capítulo 4 - Classificação

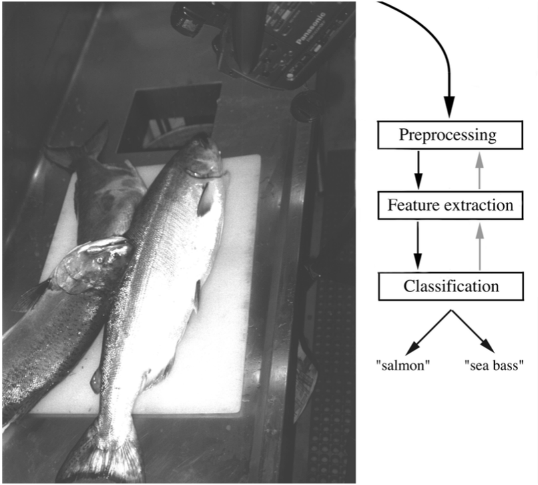
Cristiano Leite de Castro

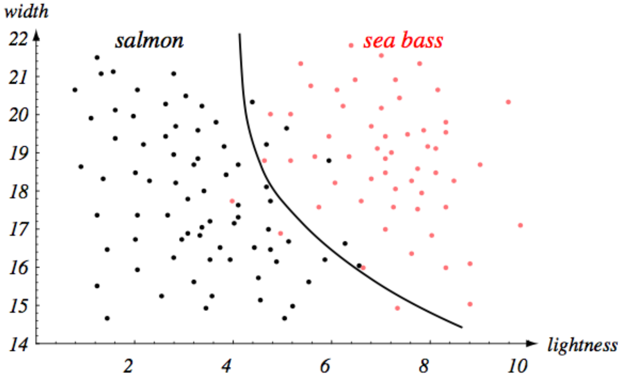
## Exemplo




É possível prever o clima a partir de fatores como, por exemplo, a temperatura?

## Exemplo


Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"


width
salmon
sea bass
22
21
20
19
18
17
16
15
14
2
4
6
8
10
lightness

- Variáveis **categóricas** assumem valores em um conjunto enumerável. ex:
  - eye color = $\{brown, blue, green\}$.
  - email = $\{spam, no\ spam\}$.
- dado um conjunto de preditores $X = \{X_1, X_2, \ldots, X_p\}$ e uma variável de resposta categórica $Y = \{1, \ldots, k, \ldots, K\}$, a tarefa de **classificação** é construir uma função $\hat{f}(x)$ que recebe como entrada o vetor $X = x$ e prediz a sua classe correspondente $Y = k$.
- lembrando que $\hat{f}(x)$ deve ser estimado a partir do conjunto de observações
$$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

- Em algumas aplicações, no entanto, pode ser mais interessante um classificador que fornece

$$P(Y = k|X = x) \ \ \forall \ k,$$

ou seja, as probabilidades de $x$ pertencer a cada uma das $K$ classes.

- Ex: faz mais sentido estimar a probabilidade de uma *insurance claim* ser uma **fraude** do que simplesmente classificá-la como: **fraude** ou **não fraude**.

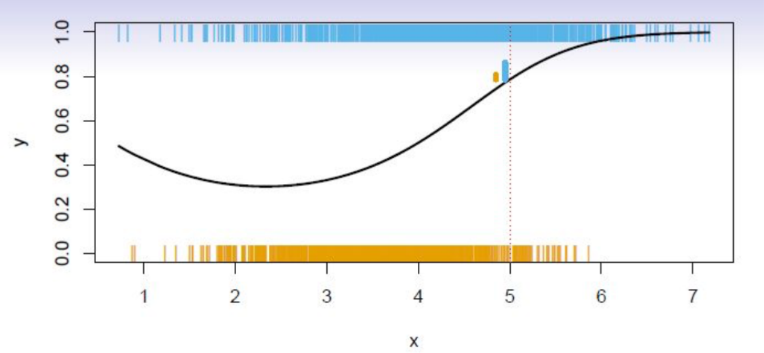
- Para variáveis categóricas binárias, tais como *tumor* $=$ {Maligno, Não Maligno} ou *email* $=$ {SPAM, NO SPAM} é comum usar o seguinte esquema de codificação

$$Y = \begin{cases} 1 & \text{se } \textbf{Sim} \\ 0 & \text{se } \textbf{Não} \end{cases}$$

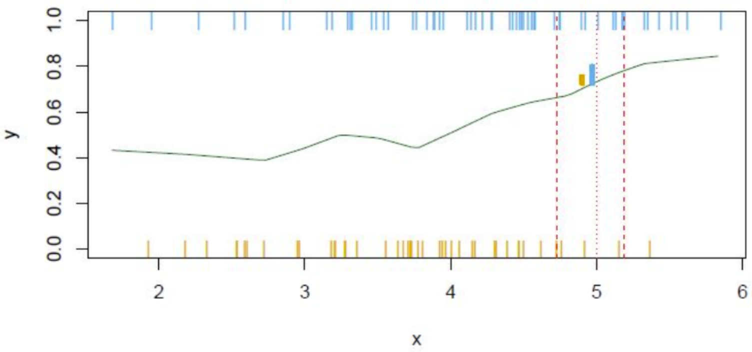
  onde $Y = 1$ é tomado como valor de referência.

- nesse caso, a tarefa de classificação se resume a modelar a função $P(Y = 1|X = x)$, abreviada como $f(x)$
- tal que $P(Y = 0|X = x) = 1 - P(Y = 1|X = x)$.

exemplo de **Função Geradora** populacional: $f(x) = P(Y = 1|X = x)$



estimando $\hat{f}(x) \approx P(Y = 1|X = x)$ por Vizinhança Local (*sliding window*)



- **Métodos Discriminativos**:
  - modelam diretamente $P(Y = k|X = x)$, abreviada como $f_k(x)$.
  - KNN, **Regressão Logística**, Redes Neurais, etc.
- **Métodos Generativos**:
  - primeiramente, modelam as densidades dos preditores por classe, i.e., $P(X = x|Y = k)$.
  - usam o Teorema de Bayes para obter $P(Y = k|X = x)$.
  - **LDA**, QDA, Naive Bayes, etc.

- **Credit Card Default Data Set**: contém info sobre 10.000 clientes
  - $Y$ : inadimplente: Sim, Não. (categórica)
  - $X_1$ : saldo mensal da fatura. (numérica)
  - $X_2$ : renda mensal. (numérica)
  - $X_3$ : estudante: Sim, Não. (categórica)
- $\hat{\pi}_{inadimplente} \approx 3\%$: proporção dos clientes inadimplentes no conjunto de dados.

- Suponha o seguinte esquema de codificação para a variável de resposta **inadimplente**

$$Y = \begin{cases} 1 & \text{se } \textbf{Sim} \\ 0 & \text{se } \textbf{Não} \end{cases}$$

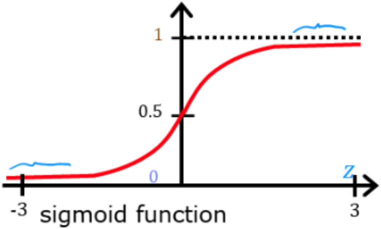
- assumindo o modelo de regressão logística, a relação entre a probabilidade de ser **inadimplente** e a variável preditora **saldo da fatura** é dada por

$$P(Y = 1|X = x) = g(\beta_0 + \beta_1 x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

onde $\beta_0$ e $\beta_1$ são os parâmetros do modelo e $g(z) = \frac{1}{1+e^{-z}}$.

## Logistic Function

Want outputs between 0 and 1



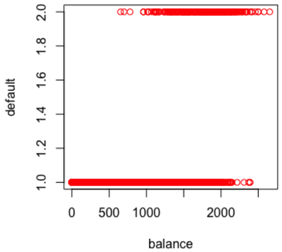

sigmoid function

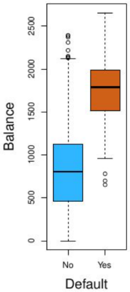
logistic function

outputs between 0 and 1

$$g(z) = \frac{1}{1+e^{-z}} \qquad 0 < g(z) < 1$$

Relação entre a variável **inadimplente** e **saldo da fatura**:




Qdo ocorre inadimplência, em geral, os clientes apresentam valores elevados de fatura de cartão de crédito (em dólares).

## Ajuste do Modelo para o Default Data Set

estimando os coeficientes $\hat{\beta}_0$ e $\hat{\beta}_1$ (estimador de Máxima Verossimilhança):

```
Call:
glm(formula = default ~ balance, family = binomial, data = Default)

Deviance Residuals:
   Min       1Q  Median       3Q      Max
-2.270  -0.146  -0.059  -0.022   3.759

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.65133    0.36116   -29.5   <2e-16 ***
balance       0.00550    0.00022    24.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600
```

os resultados mostram que existe uma relação estatisticamente significativa entre as variáveis **fatura** e **inadimplência**: $\hat{\beta}_1 = 0.0055$, com p-valor $< 2e^{-16}$.

- desde de que a relação entre $P(Y = 1|x)$ e $x$ é não linear, **não é possível** avaliar diretamente qual seria o efeito de um aumento de uma unidade na variável preditora $(x + 1)$ na probabilidade de ser inadimplente.
- a interpretação para o coeficiente $\hat{\beta}_1$ passa pela definição do conceito de **razão de chance** (*odds ratio*),

$$OR = \frac{P(Y = 1|x)}{1 - P(Y = 1|x)} = e^{\hat{\beta}_0 + \hat{\beta}_1 x}$$

- $OR$ pode assumir valores entre 0 e $\infty$.
  - $OR = 1$: chances iguais de $Y = 1$ ocorrer, i.e., equivalente a $P(Y = 1|x) = 0.5$.

- a interpretação para o efeito de $\beta_1$ é **multiplicativa** (ao invés de aditiva):

$$OR = \frac{P(Y=1|x)}{1-P(Y=1|x)} = e^{\hat{\beta}_0+\hat{\beta}_1 x}$$

- isso significa que a cada aumento de uma unidade na variável preditora $(x+1)$, a razão de chance será multiplicada por $e^{\beta_1}$

$$\frac{P(Y=1|x+1)}{1-P(Y=1|x+1)} = \frac{P(Y=1|x)}{1-P(Y=1|x)} \times e^{\hat{\beta}_1}$$

- assim, para cada aumento de uma unidade no valor da **fatura** do cartão de crédito, a razão de chance (de ser **inadimplente**) será multiplicada por

$$e^{0.0055} = 1.0055$$

  - $\hat{\beta}_1 > 0$: aumento em $X$ está associado a um aumento em $OR$ e, consequentemente, $P(Y=1|x)$.
  - $\hat{\beta}_1 < 0$: aumento em $X$ está associado a um decréscimo em $OR$ e, assim em, $P(Y=1|x)$.

- probabilidade de ser **inadimplente** para alguém com uma **fatura** de 1000.

$$p(y=1|x) = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x)}} = \frac{1}{1+e^{-(-10.6513+0.0055\times 1000)}} = 0.0057$$

- probabilidade de ser **inadimplente** para alguém com uma **fatura** de 2000.

$$p(y=1|x) = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x)}} = \frac{1}{1+e^{-(-10.6513+0.0055\times 2000)}} = 0.5863$$

- a extensão do modelo logístico para o caso multivariado ($p > 1$) é natural.
- **Default Data set**: a relação entre a probabilidade de ser *inadimplente* e as variáveis preditoras: *saldo da fatura*, *renda* e *estudante [sim]* é dada por

$$P(Y = 1|X = x) = g(\beta_0 + \beta_1 x + \beta_2 x + \beta_3 x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x + \beta_2 x + \beta_3 x)}}$$

onde $\beta_0$, $\beta_1$, $\beta_2$ e $\beta_3$ são os parâmetros do modelo e $g(z) = \frac{1}{1+e^{-z}}$.

estimando os coeficientes $\hat{\beta}_0$ , $\hat{\beta}_1$, $\hat{\beta}_2$ e $\hat{\beta}_3$:

```
Call:
glm(formula = default ~ balance + income + student, family = binomial,
    data = Default)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.469  -0.142  -0.056  -0.020   3.738

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.09e+01   4.92e-01  -22.08   <2e-16 ***
balance      5.74e-03   2.32e-04   24.74   <2e-16 ***
income       3.03e-06   8.20e-06    0.37   0.7115
studentYes  -6.47e-01   2.36e-01   -2.74   0.0062 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

os resultados mostram que existe uma relação estatisticamente significativa entre as variáveis **fatura** e **estudante [sim]** com **inadimplência**. (ver *p-values* para $\hat{\beta}_1$ e $\hat{\beta}_3$)

$\hat{\beta}_3 < 0$ indica que estudantes são **menos propensos** a serem inadimplentes que não-estudantes.

- considere um problema de classificação com $p$ preditores $X = \{X_1, X_2, \ldots, X_p\}$ e $K$ classes, isto é, $Y \in \{1, \ldots, k, \ldots, K\}$.
- a **Análise de Discriminantes** estima as probabilidades condicionais $P(Y = k|X = x)$ (indiretamente) usando a regra de Bayes

$$P(Y = k|X = x) \equiv p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

onde

- $f_k(x) \equiv P(X = x|Y = k)$: função densidade de $X$ na classe $k$.
- $\pi_k \equiv P(Y = k)$: probabilidade a priori para a classe $k$.

- Suponha o seguinte esquema de codificação para a variável de resposta **inadimplente**
$$Y = \begin{cases} 1 & \text{se } \textbf{Sim} \\ 2 & \text{se } \textbf{Não} \end{cases}$$
- A prob. de ser **inadimplente** para uma dada observação é obtida com o Teorema de Bayes
$$P(Y = 1|X = x) \equiv p(x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}$$
tal que $P(Y = 2|X = x) = 1 - P(Y = 1|X = x)$.
- Uma vez estimadas as probs. a posteriori para cada classe, a seguinte regra de decisão é adotada
$$x = \begin{cases} 1 & \text{se } P(Y = 1|X = x) > P(Y = 2|X = x) \\ 2 & \text{caso contrário.} \end{cases}$$
- Problema: como estimar $\pi_1$, $f_1(x)$, $\pi_2$, $f_2(x)$?

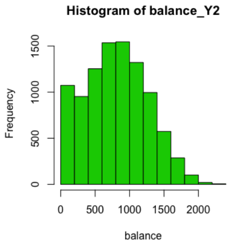
- $\pi_k$: probabilidade que uma observação amostrada aleatoriamente da população seja da classe $k$.
- Para o **Default dataset**:
  - $\pi_1 \equiv \pi_{inadimplente}$: proporção de **clientes inadimplentes** na população de interesse (Ex: $\approx$ 3%)
  - $\pi_2$: proporção de **clientes não-inadimplentes** na população de interesse (Ex: $\approx$ 97%)
- $\hat{\pi}_k$ é uma estimativa para $\pi_k$ a partir de um conjunto de observações. Geralmente calculada como

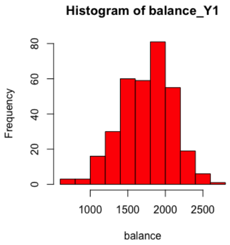
$$\hat{\pi}_k = \frac{n_k}{n}$$

onde $n_k$ é o número de observações da classe $k$ no conjunto de treinamento.

## Interpretando $\pi_k$ e $f_k(x)$ (2)

- $f_k(x) \equiv P(X = x|Y = k)$ : denota a função de densidade de $X$ para uma observação da classe $k$.
- Para o **Default dataset**: se $X$ representa a saldo da fatura do cartão de crédito, as curvas $f(x)_1$ e $f(x)_2$ representam a diferença populacional entre os valores de fatura para os inadimplentes e não-inadimplentes, respectivamente.
  - $f_1(x) \equiv P(X = x|Y = 1)$ será **pequena** se é improvável que um **cliente inadimplente** possua um valor de fatura $X = x$.
- os histogramas da variável **fatura** para as observações com $Y = 1$ (inadimplente) e $Y = 2$ podem ser consideradas estimativas para $f(x)_1$ e $f(x)_2$, respectivamente.



- **Análise de Discriminantes Lineares (LDA).**
- Premissas:
  1. $f_k(x) \sim \mathcal{N}(\mu_k, \Sigma_k^2)$ : os preditores são gerados a partir de **Gaussianas Multivariadas** com um vetor de média específico para cada classe.
  2. $\Sigma_k^2 = \Sigma^2 \, \forall \, k = 1, \ldots, K$ : as **matrizes de covariância são iguais** para cada classe.

- para um único preditor ($X_1$), tem-se que $f_k(x) \sim \mathcal{N}(\mu_k,\, \sigma_k^2)$ (Gaussiana Univariada)

$$Pr(X = x|Y = k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

- desde que $\sigma_k^2 = \sigma^2 \,\forall\, k$ então

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}$$

- **Regra de Decisão**: atribua $X = x$ à classe para a qual $p_k(x) \equiv P(Y = k|X = x)$ é máxima.

- **Função Discriminante para a classe** $k$: função linear de $X$.

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

- **regra de Decisão**: atribua $X = x$ à classe para a qual $\delta_k(x)$ é máxima.
- **superfície de decisão** para o classifcador LDA: encontre o conjunto de valores de $x$ para os quais $\delta_k(x) = \delta_l(x) \, \forall \, k \neq l$.

## Dedução da Função Discriminante (1)

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

- Como o denominador é comum a todas as classes, a classificação pode ser feita com base apenas no numerador. Assim, definimos a função discriminante $\delta_k(x)$ como o logaritmo do numerador:

$$\begin{aligned}\delta_k(x) &= \log\left[\pi_k \cdot \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)\right] \\ &= \log(\pi_k) - \frac{1}{2\sigma^2}(x-\mu_k)^2\end{aligned}$$

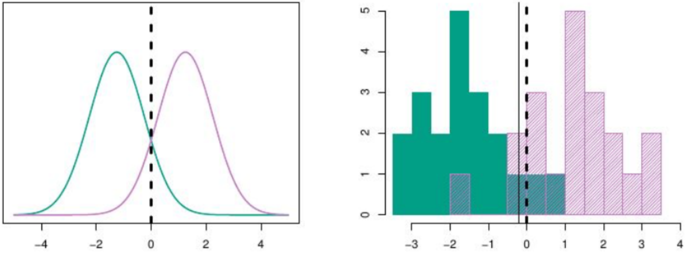
- Expandimos o quadrado:

$$
\begin{aligned}
(x - \mu_k)^2 &= x^2 - 2x\mu_k + \mu_k^2 \\
\Rightarrow \delta_k(x) &= \log(\pi_k) - \frac{1}{2\sigma^2}(x^2 - 2x\mu_k + \mu_k^2) \\
&= -\frac{1}{2\sigma^2}x^2 + \frac{\mu_k}{\sigma^2}x - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)
\end{aligned}
$$

- O termo $-\frac{1}{2\sigma^2}x^2$ é comum a todas as classes e não afeta a decisão. Assim, podemos removê-lo da função discriminante final:

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

- **Regra de Decisão**: atribua $X = x$ à classe para a qual $\delta_k(x)$ é máxima.




Example with $\mu_1 = -1.5$, $\mu_2 = 1.5$, $\pi_1 = \pi_2 = 0.5$, and $\sigma^2 = 1$.

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning*. 2013.
- Hastie, Trevor, et al. "The elements of statistical learning: data mining, inference and prediction." The Mathematical Intelligence, 2001
- Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012.
- Bishop, C. Pattern Recognition and Machine Learning. Springer-Verlag New York, Inc, 2006.