**KNN: K-Vizinhos mais próximos**

Professor Anderson 04/04/2017

X=

|  |  |  |
| --- | --- | --- |
| RENDA | CARTÃO | Y |
| 1000 | 1000 | 2 |
| 7000 | 2000 | 1 |
| 2000 | 1500 | 2 |
| 3000 | 2500 | 2 |
| 5500 | 3500 | 1 |
| 5000 | 3000 | 1 |
| 3500 | 2000 |  |
| 5500 | 2000 |  |

**1 = Feliz 2 = Triste**

O “motor” de funcionamento do KNN está na descoberta dos objetos mais próximos que (são os exemplos já conhecidos).

Para usá-los precisamos;

1. Definir um conjunto de treinamento
2. Definir uma métrica de distância
3. Definir o valor de **K**

Distância euclidiana é:

Seja “p” e “q” objetos de classificação

d(p,q) = RAIZ(n SOMATORIO i=1 (pi – qi)²)

n = número de variáveis

City-Block

N SOMATORIO i=1 |Pi-Qi|

Outras distâncias:

Hamming: Variáveis de valores 0 ou 1

Mahalanobis: Falaremos disto depois...

1. Normalizar os dados
   1. Normalização linear [0,1]
      1. X = X-XMIN/XMAX-XMIN
   2. Normalização por desvio padrão
      1. Desvio padrão é a raiz da variância (dispersão ou distância da media)
      2. Atenuação da discrepância dos dados;
      3. Considera a posição média dos valores e os graus de dispersão à média.
         1. X=(X-^-X)/desvio padrão <- Útil qnd min e max são desconhecidos

|  |  |  |
| --- | --- | --- |
| DESPESA | POR DESVIO | LINEAR |
| 1000 | -0,75 | 0,14 |
| 2000 | 0,13 | 0,43 |
| 3000 | 1,02 | 0,71 |
| 1500 | -0,31 | 0,29 |
| 1500 | -0,31 | 0,29 |
| 1000 | -0,75 | 0,14 |
| 3000 | 1,02 | 0,71 |
| 500 | -1,19 | 0 |
| 4000 | 1,90 | 1 |
| 1000 | -0,75 | 0,14 |

Media: 1550 Desvio padrão: 1131,62

Classificando um novo exemplo:

1. Calcule a distância entre o novo objeto em relação a todos os objetos de treinamento.
2. Identifique os K mais próximos
3. Identifique a classe dos K vizinhos mais próximos P/ determinar o rótulo do novo objeto

**Vantagem**

- Simples e fácil de implementar

- Apresenta resultados competitivos

**Desvantagem**

- Caro p/ bases de treino “grandes”

- Sensível à variáveis irrelevantes ou a ruídos

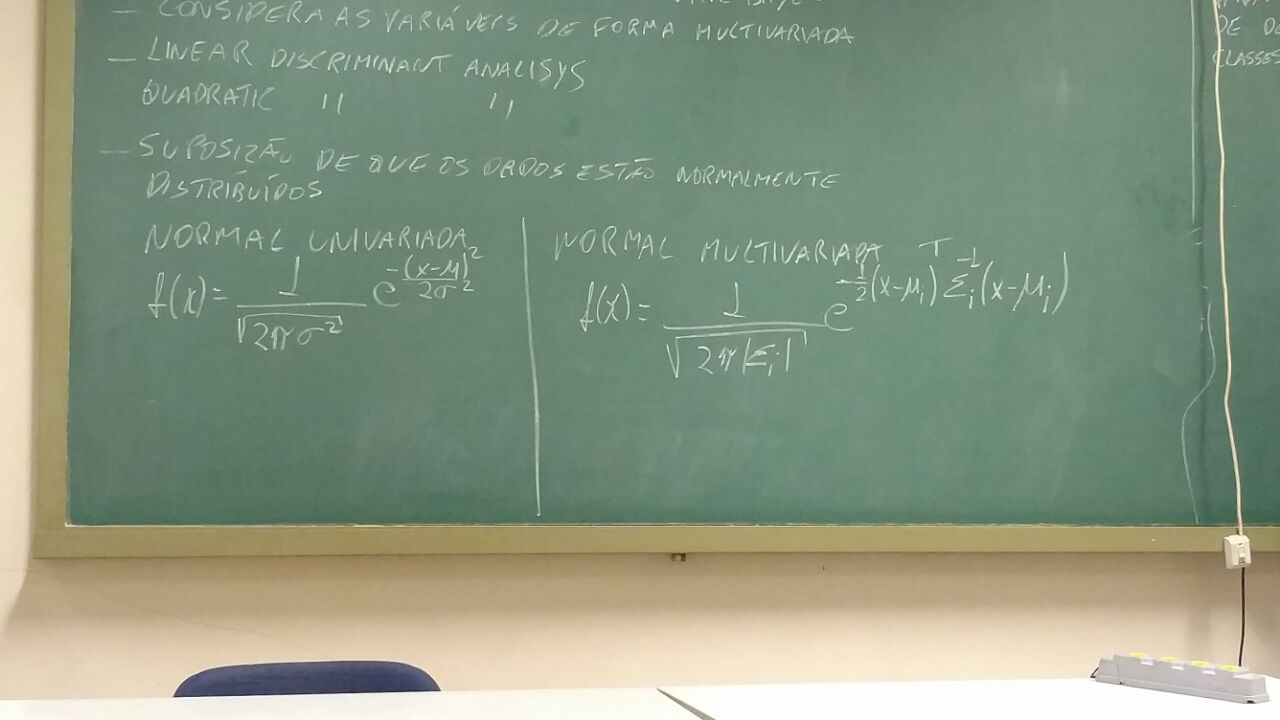
**LDA e QDA**

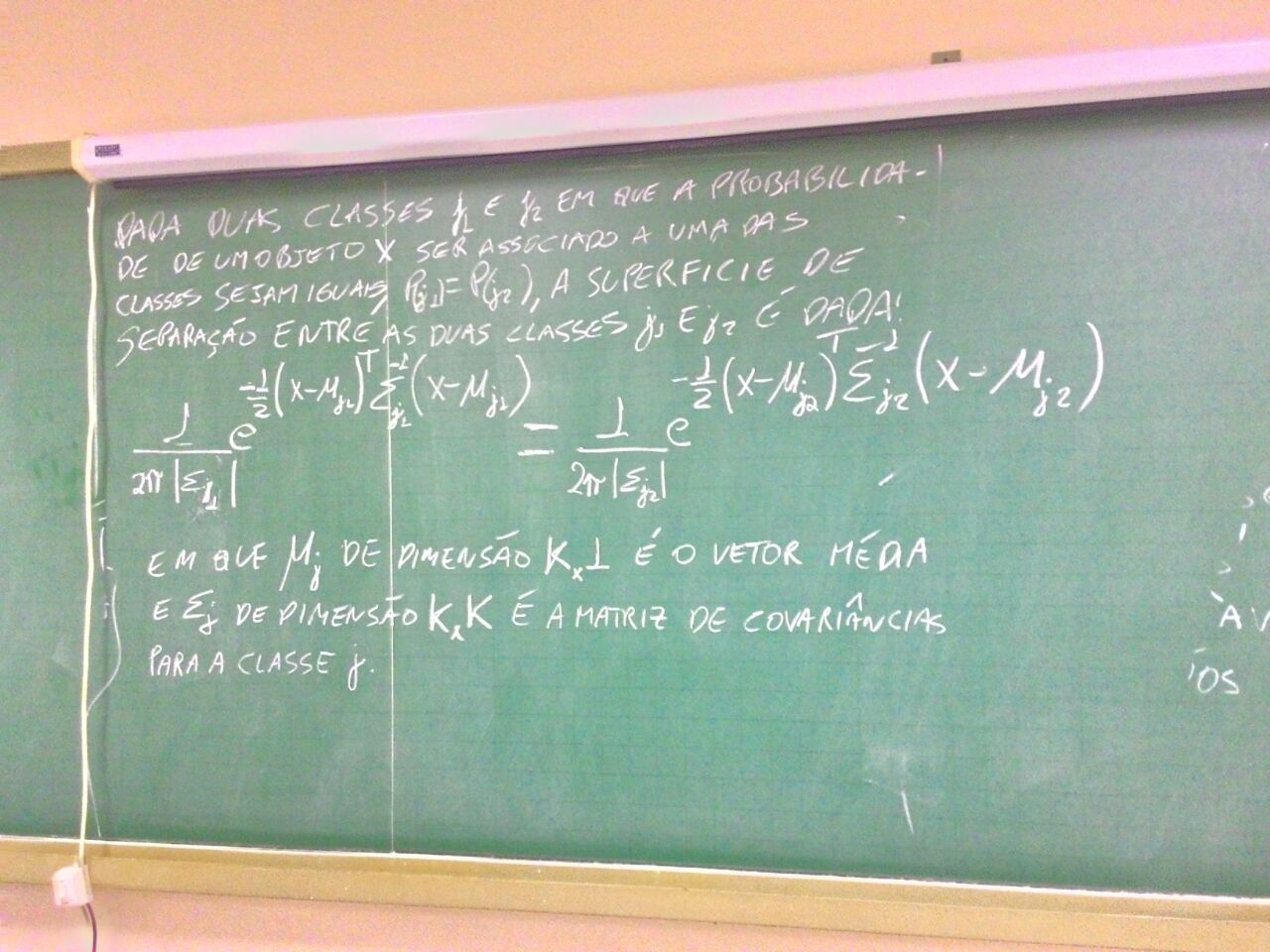
- Classificadores alternativos ao Naive Bayes.

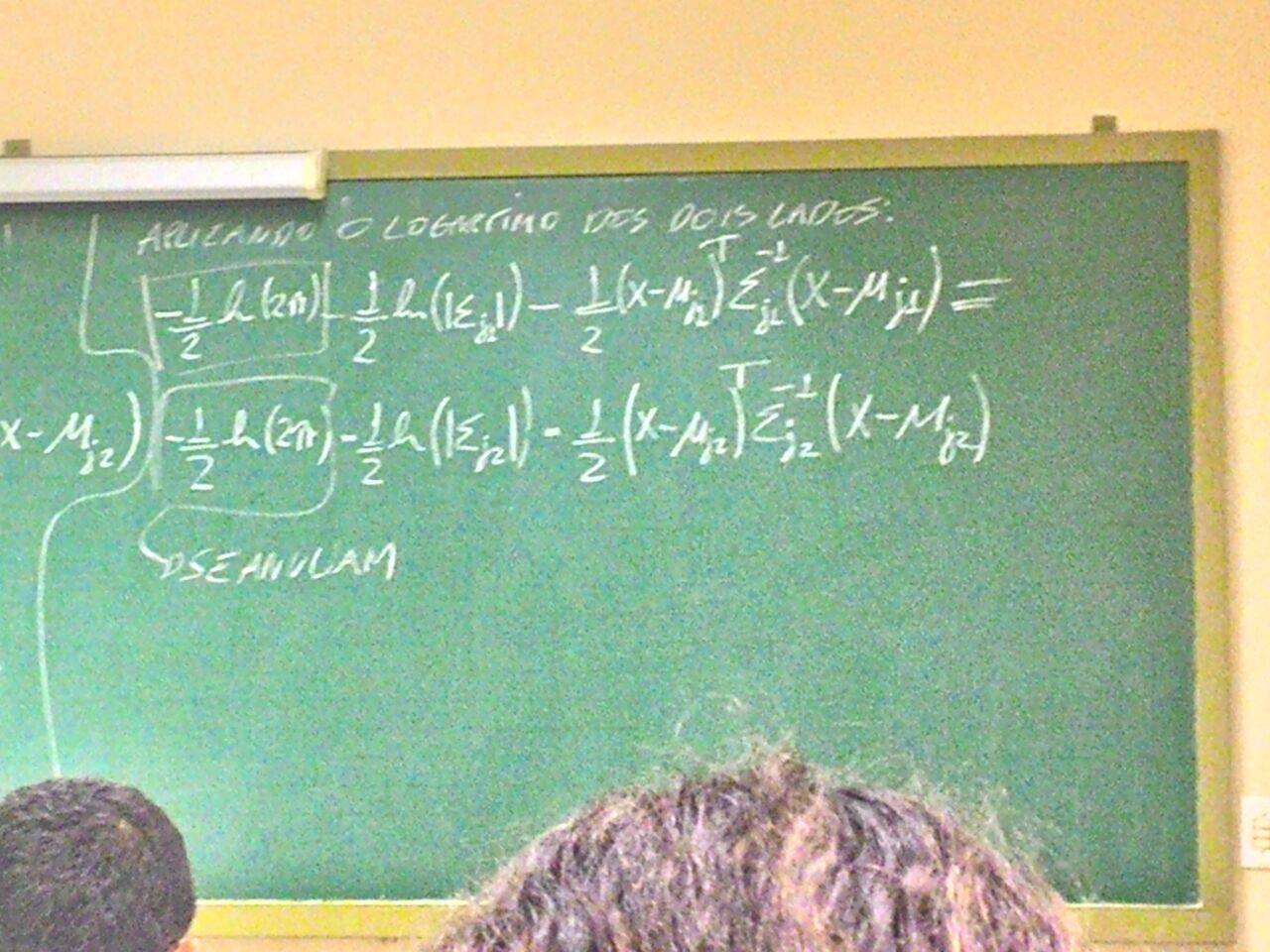
- Considera as variáveis de forma multivariada

- Linear discriminat anallisys

- Quadratic “” “”

- Suposição de que os dados estão normalmente distribuídos.

Dada duas classes J1 e J2 em que a probabilidade de um objeto X ser associado a uma das classes sejam iguais. P(J1) = P(J2). A superfície de separação entre as duas classes J1 e J2 é dada:

Aplicando o logaritmo dos dois lados:

**Algoritmo QDA:**

Dado X e Y calcule M1, E1 e M2 E2

Covariância mede até onde o limite da classe chega.

Fim treinamento

Dado um novo objeto X teste.

{{formula D1 e D2}}

**No LDA**

Calculamos uma única matriz de covariâncias para todos os objetos.