Universidade Fedeal de Minas Gerais

Exercício 5

A.P. Braga

May 9, 2017

Kernel Density Estimation

O aluno deverá implementar um classificador utilizando o KDE para a base de dados Winsconsin Breast Cancer.
Considere a equação de Bayes:
Para o cálculo das probabilidades a priori de cada classe, deve ser determinada a quantidade ocorrências desta classe no conjunto de treinamento em relação à quantidade total de amostras.

Winsconsin Breast Cancer

Deve ser utilizada a mesma base de dados do exerc icio de mistura de gaussianas, a *Winsconsin Breast Cancer* da UCI através do pacote *mlbench*.

```
> library('mlbench')
```

Lembre-se que a primeira coluna desta base de dados é um digito de identificação, ele pode ser descartado. Além disso nela existem alguns dados faltantes, assim, é sugerido que se substitua o valores não atribuídos (NA) por 0:

```
> data(BreastCancer)
> summary(BreastCancer)
> X <- data.matrix(BreastCancer[,2:10])
> X[is.na(X)] <- 0
> trainY <- as.numeric(BreastCancer$Class)
```

Novamente, como a base de dados está no formato de data frame, os comandos *data.matrix* e *as.numeric* podem ser utilizados para transformá-la em numérica.

Treinamento e Teste

A base de dados deve ser divida em dois conjuntos, um de treinamento e outro de teste, de razão a critério do aluno, preferencialmente a mesma utilizada no exercício de GMM. A metodologia para treinamento e teste para um modelo Kernel Density Estimation (KDE) é semelhante ao do Gaussian Mixture Model (GMM). Dessa forma as etapas do desenvolvimento são similares:

- Com o grupo de testes, deve ser obtido um modelo KDE para os dados de cada classe.
- O grupo de treinamento deve ser classificado de acordo com os modelos estimados no treinamento. Deve se verificar para qual densidade estimada cada dado mais se adequa.

Instruções

Neste exercício o aluno deverá:

1. Carregar os dados do pacote *mlbench* e substituir os dados faltantes por 0, por exemplo.
2. Dividir de forma aleatória os dados em grupos de treinamento de teste de acordo com uma razão pré definida.
3. Utilizar uma rotina ou função de treinamento que estime os modelos KDE.
4. Utilizar uma rotina função de teste que verifique pertinência de cada amostra a cada classe do KDE.
5. Calcular o erro quadrático médio (MSE) percentual do classificador.
6. Repetir 10 vezes os procedimentos 2 ao 5 e estimar o MSE percentual médio e o desvio padrão do classificador.

Forma de Entrega

Relatório em .doc ou .pdf, descrevendo o que foi feito, mostrando os gráficos e as informações pedidas e explicando os resultados obtidos, assim como as partes importantes do código. O relatório deve ser colocado em um arquivo .zip junto com os códigos utilizados e enviado via Moodle.