Exercícios de Aprendizado de Máquina

Data de Entrega: 23/07/2017

Não utilize funções prontas de algoritmos aprendidos em sala de aula. Implemente as suas e apresente-as na lista. Faça um relatório explicando como foi resolvido o exercício e envie junto com o código fonte.

Parte IV – Métodos de Classificação Baseados em Probabilidade

1) Para a base Car Evaluation (disponível em http://archive.ics.uci.edu/ml/), considerando que o primeiro atributo é x1, o segundo é x2 e assim por diante, estime as probabilidades:

a) P(x1 =med) e P(x2 = low)

b) P(x6=high|x3=2) e P(x2=low|x4=4)

c) P(x1=low|x2=low,X5=small) e P(x4=4|x1=med,X3=2)

d) P(x2= vhigh,X3=2|X4=2) e P(x3=4,x5=med|x1=med)

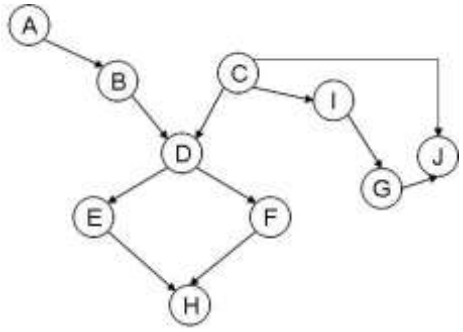
2) Aplique o *Naive* Bayes sobre a base de dados Balance Scale (disponível em http://archive.ics.uci.edu/ml/) utilizando o procedimento de *Hold-Out* dez vezes, na proporção de 75% de amostras de treino e 25% de teste. Obtenha a acurácia média e o desvio padrão da acurácia. Realize os experimentos:

a) Considerando uma distribuição Gaussiana dos atributos;

b) Discretizando os valores (em 5 partes cada atributo);

c) Discretize os valore da mesma forma que em b) usando a suavização de Laplace.

3) Para a rede bayesiana da figura abaixo, verifique as seguintes afirmações, indicando se é falso ou verdadeiro e fornecendo a devida explicação.



a) A é independe de C

b) A é independe de C tal que foi observado D

c) A é independe de C tal que foi observado H

d) F é independente de C

e) G é independente de C tal que foi observado I

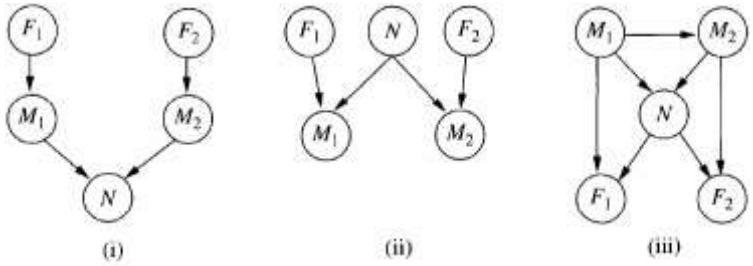
f) G é independente de C tal que foram observados I e J

g) I é independente de A tal que foi observado E

h) E e F são independentes de B tal que foi observado D

4) Dois observadores em lugares distintos obtêm medidas diferentes M_1 e M_2 do número de estrelas N em uma pequena região do céu, usando telescópios. Cada telescópio pode (com uma pequena probabilidade f) estar fora de foco (eventos F_1 e F_2), e nesse caso o astrônomo deixará de contar 3 estrelas ou mais (se N for menor ou igual a 3 ele não contará nenhuma estrela).

a) Quais das redes abaixo são representações corretas da informação acima?



b) Qual é a melhor rede? Explique.

Parte V – Métodos de Classificação Baseados em Procura

5) Para a base Car Evaluation (disponível em http://archive.ics.uci.edu/ml/):

a) Construa uma árvore de decisão com dois níveis de nó de decisão (isto é, o primeiro nó de decisão (primeiro nível), os nós de decisão abaixo dele (segundo nível) e em seguida os nós folha) usando a medida de Ganho de Informação. Selecione aleatoriamente 75% dos dados para treinamento que serão usados para construir a árvore. Retorne a estrutura da árvore construída.

b) Use os restantes 25% dos dados para avaliação. Retorne a acurácia obtida.

c) Tente obter as regras de decisão a partir da árvore construída.

6) Para a base Servo (disponível em http://archive.ics.uci.edu/ml/):

a) Construa uma árvore de regressão com dois níveis de nó de decisão (isto é, o primeiro nó de decisão (primeiro nível), os nós de decisão abaixo dele (segundo nível) e em seguida os nós folha) usando a medida de redução de desvio padrão. Selecione aleatoriamente 75% dos dados para treinamento que serão usados para construir a árvore. Retorne a estrutura da árvore construída.

b) Use os restantes 25% dos dados para avaliação. Retorne as medidas MAPE e RMSE.

c) Tente obter as regras de decisão a partir da árvore construída.

Parte V – Métodos de Agrupamento

7) Para as bases de dados Spiral e Jain (disponíveis em http://cs.joensuu.fi/sipu/datasets/), agrupe os dados em 3 e 2 grupos, respectivamente, usando kmeans e clusterização hierárquica. Avalie os resultados com a métrica de pureza, que é calculada de forma semelhante a acurácia: para cada cluster verifique qual foi a classe predominante, amostras pertencentes a outras classes estão no grupo errado. O resultado da métrica é o número total de amostras predominantes nos clusters dividido pelo número total de amostras, multiplicado por 100%. Calcule também a métrica distância intra-inter clusters, mostrada abaixo. Faça os experimentos com a distância Euclidiana. Gere gráficos com os grupos formados pelo kmeans e clusterização hierárquica. Comente os resultados e os métodos usados. Lembre-se de não usar o atributo da classe para agrupar os dados.

$$\boldsymbol{dist\_intra\_inter} = \frac{\boldsymbol{dist_{intra}}}{\boldsymbol{dist_{inter}}}$$

Seja N o número de amostras, K o número de clusters, n_j o número de amostras no j-ésimo cluster, μ_j o centroide do j-ésimo cluster e μ o centroide dos centroides, obtém-se as equações:

$$dist_{intra} = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{n_j}\left\|x_i^j - \mu_j\right\|^2$$

$$dist_{inter} = \frac{1}{K}\sum_{j=1}^{K}\left\|\mu_j - \mu\right\|^2$$