MAC425/5739 INTELIGÊNCIA ARTIFICIAL

TERCEIRO EXERCÍCIO-PROGRAMA
APRENDIZAGEM SUPERVISIONADA

PROF. DENIS DERATANI MAUÁ

Resumo. Nesta atividade você testará seus conhecimentos sobre aprendizagem supervisionada num problema real de classificação de mensagens de texto usando os classificadores Nearest Neighbors, Árvores de Decisão e Naive Bayes.

1. Considerações iniciais

Você deve entregar um relatório conciso (máx. 5 páginas) em formato pdf até às 23:55 horas do dia 22 de novembro de 2015 através do sistema PACA. Não é necessário entregar nenhum código-fonte ou arquivo de dados. O relatório deve responder a todas as questões enunciadas a seguir. Não se esqueça de colocar seu nome e número USP no cabeçalho. Procure fornecer justificativas objetivas, preferencialmente suportadas por dados (gráficos e tabelas ilustrando argumentos são bem-vindos). Você pode utilizar qualquer software, linguagem de programação e biblioteca que desejar para resolver este exercício (mas lembre-se de documentar todas suas escolhas).

2. Pré-processamento

Para este exercício, usaremos o conjunto de dados `smsspam.txt.zip`, que acompanha este enunciado. Note que o arquivo encontra-se comprimido. O arquivo (descomprimido) contém mensagens sms reais categorizadas em spam (indesejadas) e ham (não indesejadas), como no exemplo da Figura 1. O arquivo contém uma mensagem por linha, no formato `categoria, mensagem`. Mais informações sobre estes dados podem ser obtidas em `http://www.dt.fee.unicamp.br/%7Etiago/smsspamcollection/`. Antes de construir classificadores, é preciso pré-processar o conjunto de dados, a fim de obter uma codificação dos dados como vetores de características, e separar os dados em conjuntos próprios para a estimação e avaliação dos modelos.

Sua primeira tarefa é separar as mensagens em conjuntos de treinamento e validação, cuidando para que a proporção entre spam/ham em cada conjunto seja semelhante à proporção no conjunto original. O conjunto de treinamento deve contar aproximademente 70% das mensagens. O objetivo do conjunto de validação é simular o desempenho do seu sistema de classificação em dados novos, testando assim o poder de generalização do classificador. Portanto, o conjunto de validação deve ser utlizado apenas na última etapa para comparar os classificadores obtidos.

Após segmentar o conjunto de dados, você deve transformar cada mensagem em um vetor de características de dimensão fixa. Para isso você deve ler o capítulo

ham	Hope you are having a good week. Just checking in
ham	K..give back my thanks.
ham	Am also doing in cbe only. But have to pay.
spam	"complimentary 4 STAR Ibiza Holiday or L10,000 cash needs your URGENT collection. 09066364349 NOW from Landline not to lose out! Box434SK38WP150PPM18+"
spam	okmail: Dear Dave this is your final notice to collect your 4* Tenerife Holiday or #5000 CASH award! Call 09061743806 from landline. TCs SAE Box326 CW25WX 150ppm

FIGURA 1. Exemplo de mensagens sms no conjunto de dados.

22.2 do livro-texto (AIMA). Recomenda-se usar a representação de saco de palavras (bag-of-words), removendo palavras e expressões muito comuns (stopwords) e também as pouco frequentes. Quaisquer estatísticas devem ser computadas utilizando apenas o conjunto de treinamento (embora qualquer transformação dos dados deve ser realizada também no conjunto de validação). Você deve testar várias codificações do problema e compará-las (por exemplo, usando contagens de palavras, frequência relativa no documento, presença/ausência etc.). Em geral, a melhor codificação para um determinado classificador não é a melhor codificação para outro.

3. Nearest Neighbors

O primeiro classificador a ser testado é o classificador por vizinhança. Os parâmetros desse classificador são a medida de distância (Hamming, Manhattan, euclideana etc.) e o número de vizinhos. As diferentes configurações de parâmetros devem ser comparadas utilizando validação cruzada no conjunto de treinamento. Você deve avaliar cada escolha de configuração dos parâmetros em diferentes codificações do problema.

3.1. **Questões.**

- Forneça uma estimativa para o erro médio de predição (com desvio-padrão) em novos dados para cada configuração/codificação testada.
- Qual a melhor configuração de parâmetros?

4. Árvores de Decisão

Em seguida, você deve aprender um classificador baseado em árvores de decisão. Você deve testar diferentes critérios de seleção de atributos (Entropia, índice Gini e erro de classificação) e de parada, e selecionar a melhor combinação (utilizando validação cruzada no conjunto de treinamento).

4.1. **Questões.**

- Forneça uma estimativa para o erro médio de predição (com desvio-padrão) em novos dados para cada configuração/codificação testada.
- Qual a melhor configuração de parâmetros?

5. Naive Bayes

O último classificador a ser testado é o classificador probabilístico ingênuo (Naive Bayes). Note que esse classificador requer que as variáveis de atributos sejam categóricas (ou seja, com um número finito de valores). Dependendo da condificação dos vetores de atributos você precisará discretizar os dados antes de classificá-los. Uma maneira de discretizar um atributo é dividir seu domínio em intervalos equidistantes (de mesmo tamanho). Outra maneira mais eficaz é construir uma árvore de decisão contendo apenas aquele atributo. Você deve testar o classificador usando suavização de La Place nos parâmetros e escolher a melhor configuração (utilizando apenas o conjunto de treinamento).

5.1. **Questões.**

- Como os dados foram discretizados?
- Forneça uma estimativa para o erro médio de predição (com desvio-padrão) em novos dados para cada configuração/codificação testada.
- Qual o melhor valor para o parâmetro de suavização?

6. Validação

Você deve comparar o desempenho dos melhores classificadores (com seus parâmetros estimados usando o conjunto de treinamento) no conjunto de validação.

6.1. **Questões.**

1. Qual classificador é melhor no conjunto de validação?
2. A comparação entre os desempenhos dos classificadores muda qualitativamente quando feita no conjunto de treinamento (com validação cruzada) e quando feita no conjunto de testes? Especule a razão.
3. Quais as possíveis razões da superioridade do melhor classificador?
4. A comparação é confiável? Argumente sua resposta em caso afirmativo, ou descreva uma metodologia que a deixe mais confiável em caso de resposta negativa.