Statistical Machine Learning

Lista 3

Rafael Izbicki

Lista em trios. Apenas um do grupo deve entregar a lista. Por favor entregue a lista em PDF.

NÃO COPIE!

Exercício 1. Baixe o conjunto de dados https://www.dropbox.com/s/ud93k0b122awvgp/voice.csv.zip. Cada observação deste banco é relativa a uma fala de um indivíduo. As covariáveis indicam características do som emitido por ele; a variável resposta indica se ele é do sexo masculino ou feminino.

Seu objetivo é criar classificadores para predizer a variável resposta com base nas covariáveis disponíveis. Para tanto, você deverá implementar os seguintes classificadores, assim como estimar seus riscos via conjunto de teste:

- Árvore de Classificação. Mostre a árvore gerada antes e depois da poda.
- Regressão Logística. Mostre os coeficientes estimados.
- Naive Bayes.
- KNN.
- Mais um classificador de sua escolha.

Responda ainda as seguintes perguntas:

- Qual o melhor classificador segundo o risco estimado? Discuta.
- Para os classificadores baseados em estimativas de probabilidade, faça também as curvas ROC com o conjunto de teste. Faça também a tabela de confusão quando o corte usado é 0.5 e também quando o corte é aquele que maximiza sensibilidade mais especificidade. Comente.

Exercício 2. (*Pós-graduação*) Considere novamente o banco de dados do exercício 1. Iremos agora simular uma situação em que há covariate shift nos dados. Para isso, eu dividi esse conjunto em dois:

- Labeled: https://www.dropbox.com/s/orhgwbhb70h70u6/voice_labeled.csv?dl=0
- Unlabeled: https://www.dropbox.com/s/gfzzxpfjbfc1zu5/voice_unlabeled.csv?dl=0

Essa divisão foi feita de forma criar covariate shift artificalmente. O conjunto labeled faz o papel do conjunto rotulado que teriamos em mãos. O conjunto unlabeled faz o papel do conjunto não rotulado que teriamos em mãos. Ou seja, na prática, não conheceríamos os rótulos desse conjunto. Vamos utilizá-los aqui apenas para verificar o desempenho dos diferentes métodos.

- Com alguns gráficos, você consegue encontrar algumas covariáveis para as quais há diferença de distribuição nos dois grupos?
- Ajuste uma regressão logística (com penalização L1) e um KNN ao conjunto labeled (com escolhas de tuning parameters usando apenas esse conjunto). Calcule seu desempenho preditivo no conjunto unlabeled
- Estime os pesos de correção usando uma regressão logística com penalização.
- Reajuste a regressão logística original usando os pesos encontrados no item anterior. Calcule seu desempenho preditivo no conjunto unlabeled. Qual dos três métodos ajustados forneceu melhores resultados?

Exercício 3. Este é um exercício bastante aberto. Você terá espaço para fazer suas escolhas conforme achar mais apropriado. Este exercício é desafiador; comece cedo!

A base de dados deste exercício foi extraída do IMBD. Ela contém títulos e resumos de filmes. Seu objetivo é definir quais *tags* foram atribuídas a cada filme. Há um total de 82 *tags* (por exemplo, cult, horror, gothic, murder etc). O conjunto de treinamento está em https://www.dropbox.com/s/6zo01p8cdrx4e7f/labeled_data.csv?dl=0. Os resultados deste exercício serão avaliados de duas formas:

1. Abaixo estão algumas direções sobre o que deve ser feito. Sua nota para esta lista será baseada no cumprimento destes quisitos.
2. Você também deverá submeter suas predições para as observações do banco https://www.dropbox.com/s/ekkykll6je6gl22/unlabeled_data.csv?dl=0, que não tem respostas disponíveis, para o google classroom. Eu irei então verificar a performance de suas predições com relação às tags usadas na realidade. A função de perda utilizada será a média da estatística F1 em cada instância. Em outras palavras: para cada instância, faremos a média harmônica entre a precisão (a fração de tags de fato usadas entre as que foram elencadas como estando presentes) e o recall (a fração de tags elencadas como presentes entre as que de fato foram usadas) (veja https://en.wikipedia.org/wiki/Precision_and_recall). O score final será dado pela média da estatística F1 em todas as instâncias do conjunto de teste.

   O grupo que tiver melhores predições obterá **um ponto a mais na média**. O grupo que for o segundo colocado obterá **meio ponto a mais na média**. **Atenção**: seu arquivo deve ter EXATAMENTE o formato do banco não rotulado enviado aqui , mas trocando os NA's por 0's e 1's conforme suas predições.

   Este arquivo deverá ter a extensão .csv e ter como nome o nome dos integrantes do trio concatenados. Por exemplo: `AndreMariaJoao.txt`. **A predição deve ser feita exclusivamente com base no conjunto de treinamento! Isto é, não é permitido usar informações externas.**

Direcionamento: Você deve:

- ajustar no mínimo 10 modelos de classificação
- testar ao menos duas formas de criar as covariáveis
- comparar os modelos propostos com a métrica que você julgar razoável, indicando qual obteve melhor desempenho.

Note que **os diferentes modelos não são necessariamente provindos de métodos diferentes.** Por exemplo, você pode ajustar uma regressão logística com uma forma de criar covariáveis, e uma segunda com outra

forma. Isso conta como 2 modelos. Você também não precisa usar o conjunto de dados inteiro, mas certamente isso irá te ajudar.

Use sua criatividade!