Mineração de Dados / Statistical Machine Learning

Lista 2

Rafael Izbicki

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

NÃO COPIE!

Exercício 1. (*Pós-graduação*) Implemente uma função no R que recebe como argumentos x.train (univariado), y.train, x.test e h e que retorna como solução o estimador da regressão linear local com kernel gaussiano para cada elemento do conjunto de teste. Mostre um exemplo numérico com dados simulados (digrama de dispersão) e a regressão estimada para diferentes valores de h. Adicione também a regressão real nele.

Exercício 2. Neste exercício você irá implementar algumas técnicas vistas em aula para um banco de dados de filmes (https://www.dropbox.com/s/6ltw600uoiynd3t/TMDb_updated.CSV.zip?dl=0). O objetivo aqui é conseguir criar uma função que consiga predizer para onde uma a nota média de um filme (vote_average) com base em seu resumo (overview).

(a) Leia o banco e faça o processamento necessário para transformar os textos em uma matriz documento-termo. Divida o conjunto fornecido em treinamento, validação e teste, justificando as porcentagens escolhidas para cada grupo. Utilizaremos o conjunto de treinamento e validação para ajustar os modelos. O conjunto de teste será utilizado para testar seu desempenho.

(b) Utilize validação cruzada (*data splitting*) para escolher o melhor k. Plote k vs Risco estimado. Utilizando o conjunto de teste, estime o risco (e seu erro padrão) do KNN para o melhor k.

(c) Ajuste uma regressão linear para os dados usando o conjunto de treinamento mais o de validação via lasso (lembre-se que a função que ajusta o lasso no R já faz validação cruzada automaticamente: ao contrário do KNN, neste caso não é necessário separar os dados em treinamento e validação). Qual o lambda escolhido? Plote lambda vs Risco estimado. Quais foram as variáveis mais importantes no ajuste (mostre os coeficientes estimados)? Utilizando o conjunto de teste, estime o risco (e seu erro padrão) do lasso para o melhor lambda.

(d) Ajuste uma floresta aleatória para os dados usando o conjunto de treinamento mais o de validação (lembre-se que praticamente não há tuning em florestas). Estime o risco com o teste. Faça um gráfico de importância de covariáveis. Como as importâncias se comparam com as obtidas pelo estimador linear?

(e) Ajuste boosting com xgboost, catboost ou lightboost, com número de iterações escolhidos via early-stopping no conjunto de validação. Estime o risco com o teste. Faça um gráfico de importância de covariáveis. Como as importâncias se comparam com as obtidas pelo estimador linear e pela floresta?

(f) Ajuste uma rede neural com early-stopping. Estime o risco com o teste. (*Pós-graduação*) Adicione drop-out na rede. O desempenho melhora?

(g) (*Pós-graduação*) Ajuste um kernel ridge regression com λ escolhido utilizando o conjunto de validação. Estime o risco (e seu erro padrão) com o teste.

(h) Plote os valores preditos versus os valores observados para o conjunto de teste em cada um dos métodos. Inclua a reta identidade.

(i) Qual modelo teve melhores resultados? Leve em conta os erros-padrão nessa análise.