UNIVERSIDADE FEDERAL DO CEARÁ

Centro de Tecnologia
Departamento de Teleinformática

REVISÃO

Aluno: Arthur Baquit Reis
Matrícula: 385185

Disciplina: Inteligência Computacional Aplicada
Professora: Michela Mulas
Data: 12 de dezembro de 2017

Fortaleza, Ceará
2017

Sumário

Perguntas e Respostas - Revisão de ICA

1 General concepts

1. "Predictive modelling is the process by which a model is created or chosen to try to best predict the probability of an outcome". Explain the key ingredients (such as outcome, predictors, training and test sets, etc.) and steps of the predictive modelling process.

   Resposta:

   (a) Os principais componentes podem ser descritos como:
   - Preditores: São nossas variáveis, ou seja, as entradas do sistema que estarão variando e estaremos analisando como estas variações alteram o sistema
   - Saída: O nome já diz: é a saída do sistema. Aqui nós usaremos como parâmetro para montar nosso modelo.
   - Conjunto de Dados: Aqui nós temos nossos conjuntos de dados, ou seja, nossas observações. A partir dele, temos o Training Set e o Test Set. O primeiro serve para usarmos a fim de montar nosso modelo, enquanto o segundo serve para ver o quão bem nosso modelo está funcionando.

   (b) Os nossos passos geralmente seguem o seguinte modelo:
   i. Análise do Conjunto de Dados: Aqui nós iremos dar uma olhada no nosso conjunto de dados para pensar na abordagem que iremos fazer no pré-processamento e começar a pensar em qual modelo iremos utilizar para resolver nosso problema.
   ii. Pré-Processamento: Após analisar os dados, iremos, se necessário, fazer alguma alteração nele a fim de adequar nossos dados ao modelo escolhido.
   iii. Escolha e implementação do modelo: Após fazer o pré-processamento, iremos modelar nosso predictive model. Aqui escolheremos se teremos uma abordagem linear, não linear, se iremos adotar métodos de redução de dimensão (como o PLS, por exemplo), ou se nosso conjunto de dados já tem poucos preditores e essa abordagem não é útil, etc.
   iv. Avaliação do modelo: Iremos avaliar o quão bem nosso modelo está agindo, utilizando nosso conjunto de treino. Para isso, iremos ver como estão os parâmetros R-squared, RMSE, curva de ROC, etc.

2. Explain whether each scenario is a classification or regression problem and indicate whether we are most interested in inference or prediction. Provide n and p.

   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

Resposta: Como estamos trabalhando com uma saída contínua (salário), escolheremos um modelo de **regressão**. Como são 500 firmas, teremos **n = 500**. Também teremos três preditores, visto que a quarta coluna representa nossa saída. Deste modo, teremos **p = 3**.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Resposta: Aqui iremos trabalhar com 'sucesso' ou 'falha', ou seja, **classificação**. Foi-se coletado dados de 20 produtos similares, então **n = 20**. De cada produto, foram recolhidos 14 dados, sendo uma delas nossa saída. Então teremos **p = 13**.

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Resposta: Outro caso de saída contínua, então caso de **regressão**. Os dados foram recolhidos semanalmente durante o ano de 2012, então temos **n = 52** observações. Em cada observação, foram-se coletadas 4 informações, sendo uma delas nossa saída. Deste modo **p = 3**.

3. Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Resposta:

- Se uma pessoa voltaria para o restaurante. A saída seria 'sim' ou 'não'. Os preditores poderiam ser: tempo para ser atendido (em segundos), tempo para o pedido ficar pronto (em segundos), preço da conta por pessoa (em Reais), sabor da comida (variável de 1 à 5, sendo 5 excelente.) e satisfação quanto ao atendimento (variável também de 1 à 5, sendo 5 excelente).
- Se um determinado investimento vale a pena ser investido. A saída seria 'Vale a pena' e 'Não vale a pena'. Os preditores seriam: retorno (em porcentagem), risco (1 para alto risco e 0 para baixo risco), duração (1 para curta duração e 0 para longa duração), livre de impostos (1 para sim, 0 para não) e valor investido (em reais).
- Se o solo está saudável ou não. A saída seria 'Saudável' e 'Não Saudável'. Os preditores poderiam ser: umidade do solo (em porcentagem), temperatura e alguns indicadores de minerais (como cálcio, magnésio, nitrogênio, fósforo). Deste modo

4. Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

   Resposta:

   - a
   - a
   - a

2 Predictive models

5. Data preprocessing techniques generally refer to the addition, deletion or transformation of training data set. Explain why data preprocessing is needed and what techniques can be used for the purpose.

   Resposta: O pré-processamento é importante por alguns motivos, dentre eles, fazer o nosso data set se adequar aos modelos que iremos inserir. Como podemos ter preditores em porcentagem, por exemplo, e outros que medem a massa de um determinado objeto em gramas, teremos uma grande divergência na escala dos nossos dados, o que causaria nosso modelo a dar mais peso para um determinado preditor baseado somente em sua escala. Esse exemplo é um de muitos que mostra a importância de fazer algumas transformações nos nossos dados, como normalização, escalamento, box-cox e centralização. Também podemos retirar alguns dados, pois podemos querer retirar dados que são muito correlacionados - já que existem modelos que podem falhar com altos preditores correlacionados. Podemos, também, querer fazer um PCA nos nossos dados para diminuir a dimensão do nosso problema. Ou ainda, podemos retirar alguns preditores que achamos que não contem informações boas, mas adicionar outro - conhecidos como 'dummy variables'. Enfim, essas transformações garantem que nós não iremos possuir dados irrelevantes nem mal-distribuidos, para assim garantir que nosso modelo tenha seu melhor desempenho.

6. Suppose that you are given a set of data consisting of n observations, p predictor variables and l classes.

   (a) How would you perform the unconditional mono-variate analysis of each of the d predictors?

   Resposta: Iria performar o histograma, calcular a skewness, a média e a variância de cada um. Isso me ajudaria a ter uma noção de como os nossos dados estão se comportando e quais as possíveis transformações seriam interessantes de usar neles.

   (b) How would you perform a class-conditional mono-variate analysis of each of the predictors?

Resposta: Plotaria o histograma de cada variável separado por classe. Aqui é interessante para tentar ver se existe algum preditor consegue, sozinho, separar uma classe - ou seja, possuindo distribuição distinta em cada classe.

(c) What kind of results would you expect from this analysis?

Reposta: Já falei um pouco em cada item, mas acho importante também ressaltar o uso do bi-variate, ou seja, fazer a análise entre os preditores. Acho que seria útil plotar o scatter-plot dos preditores e procurar algum comportamento linear entre eles. Também seria útil plotar o heatmap da matriz de correlação.

7. Principal component analysis is a commonly used method for data transformation for multiple predictors.

(a) What are the objectives of Principal Components Analysis (PCA)?

Resposta: O principal objetivo é reduzir a dimensão do conjunto de dados. Para fazer isso, ele procura os preditores que possuem maior autovalores e faz a rotação do eixo baseado em seus autovetores associados usando a matriz de covariância. Importante ressaltar que ter o maior autovalor é equivalente a ser o que tem a maior similaridade com o resto dos preditores. Desde modo, podemos fazer a rotação dos eixos, excluindo os demais preditores e conseguimos reduzir a dimensão sem perder muita informação.

(b) What type of data should be used for PCA?

Resposta: Numérico.

(c) How many principal components should be retained?

Resposta: Isso vai depender dos nossos dados. Geralmente fazemos uma análise de quão fiel é os nossos dados transformados em relação ao dataset atual. Deste modo, retiramos mantendo uma faixa de pelo menos 90%, 95%.

8. Explain the meaning of cross-validation and how the k-fold cross validation is implemented.

Reposta: A intenção do cross validation é usar o mesmo conjunto de dados de treino para fazer a validação do seu modelo. Ele funciona separando o nosso conjunto de dados em 'k' conjuntos menores. Um desses conjuntos é usado para treinar e enquanto os outros k-1 são usados para validar o modelo. Deste modo, teremos um data-set de treino de tamanho n/k, enquanto o de validação terá (k-1)n/k. Ele faz isso com todos os subconjuntos e escolhe o que teve melhor desempenho para ser nosso modelo final.

9. What are the advantages and disadvantages of the k-fold CV relatively to

(a) The validation set approach?

Resposta: As vantagens do CV são que: ele possui uma ideia simples e é facilmente implementável. Entretanto, por separar o conjunto de dados em poucos tamanhos (geralmente usa-se k=5 ou k=10), o erro estimado gerado por ele tende a superestimar o erro do test set no model fit.

(b) LOOCV?

Resposta: A vangatem é que podemos gerar um modelo muito mais preciso, pois estamos testando caso a caso. As desvantagens desse caso particular de CV são que tem um custo operacional muito grande e que pode gerar over-fit.

10. Discuss and explain the "bias-variance" tradeoff.

Reposta: Bom, primeiro é importante falar dos dois tipos de erros, o devido ao bias e o devido a variância. O primeiro seria o erro 'bruto', ou seja, a diferença entre a saída predita pela saída esperada. Já o segundo seria algo parecido com 'como um dado ponto específico, ao repetir a predição várias vezes vai variar'. Assim, podemos colocar quatro situações, ilustradas pela Figura 1.

3 Regression methods

11. Suppose we have a data set with five predictors, $X_1 = GPA$,$X_2 = IQ$,$X_3 =$ Gender (1 for Female and 0 for Male),$X_4 =$ Interaction between GPA and IQ and $X_5 =$ Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$,$\hat{\beta}_2 = 0.07$,$\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$,$\hat{\beta}_5 = -10$.

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.

ii. For a fixed value of IQ and GPA, males earn more on average than females.

iii. or a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

Resposta: Antes de começar a modelar a reta, iremos supor que a relação 'interação' descrita no texto é modelada pela simples multiplicação entre os termos. Assim, podemos ver que a reta linear fica dividida em dois casos, a Equação 1 para o homem e Equação 2 para as mulheres. Deste modo,

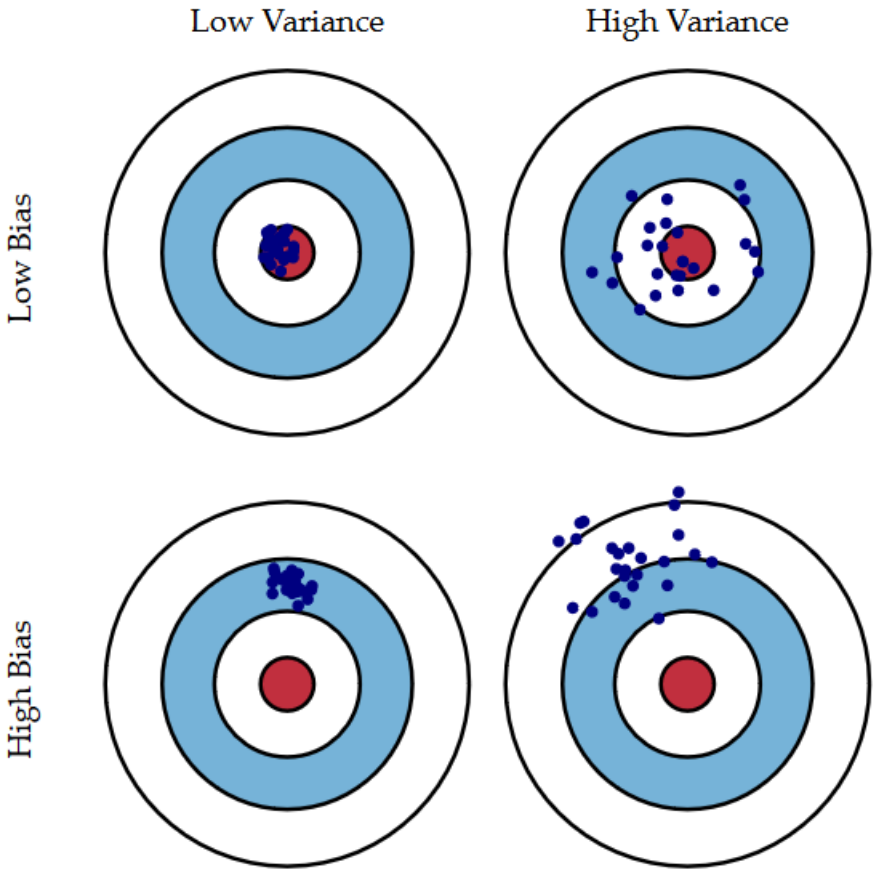

Figura 1: Figura ilustrando "Bias-Variance"

podemos notar que se GPA > 3.5 o salário das mulheres será inferior ao dos homens. Deste modo, a resposta correta é o **item iii**.

$$y = 50 + 20\text{GPA} + 0.07\text{IQ} + 0.01\text{GPA} * \text{IQ} \tag{1}$$

$$y = 50 + 20\text{GPA} + 0.07\text{IQ} + 35 + 0.01\text{GPA} * \text{IQ} - 10\text{GPA} \tag{2}$$

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0

Resposta: Basta substituir na Equação 2. Deste modo teremos que o salário estimado será de 137.1k dólares o ano.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Resposta: Falso. O que nos diz o efeito de interação entre os preditores, seria a covariância entre eles, e não o β.

12. The objective of simple linear regression is to find the plane that minimizes the sum-of-squares errors between the observed and the predicted response.

(a) How can we write this linear relationship between a quantitative response Y and the predictors X?

Resposta: Da forma de uma reta, $y = \beta_0 + X_1\beta_1 + \cdots + X_P\beta_P$

(b) How can we estimate the coefficients of a simple linear regression model?

Resposta: O parâmetro β pode ser estimado de acordo com a Equação 3.

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y \tag{3}$$

Sendo

- Y é matriz de dimensão $n \times 1$;
- X é uma matriz de dimensão $n \times p + 1$, sendo na i-ésima linha da forma: $(1, x_1, \ldots, x_p)$;
- β é a matriz de dimensão $p + 1 \times 1$

Para provar a fórmula, primeiro vamos achar o SSE, que é a função que gostaríamos de minimizar:

$$\begin{aligned} SSE(\beta) &= \Sigma(y_i + \hat{y}_i)^2 = (y - X\beta)^T(y - X\beta) \\ &= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \\ &= y^T y - 2y^T X\beta + \beta^T X^T X\beta. \end{aligned}$$

Sabemos também, que na diferenciação matricial, temos:

$$\frac{d}{d\beta}\left\{y^T y\right\} = 0 \qquad \frac{d}{d\beta}\left\{y^T X\beta\right\} = y^T X \qquad \frac{d}{d\beta}\left\{\beta^T X^T X\beta\right\} = 2X^T X\beta \tag{4}$$

Deste modo, é evidente que:

$$\frac{dSSE(\beta)}{d\beta} = -y^T X + X^T X\beta = 0 \tag{5}$$

Por fim, temos que os parâmetros beta que minimiza o erro é:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y \tag{6}$$

(c) How can we assess the accuracy of the coefficient estimate?

Resposta: Através do R-squared e o RMSE. O primeiro mede a variância, de acordo com a Equação 7, enquanto o RMSE mede o erro absoluto, de acordo com a Equação 9.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

Também podemos escrever de outro modo:

$$R^2 = 1 - \frac{\text{SSE}}{\text{TSS}} \tag{8}$$

Equação do RMSE:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{9}$$

(d) Discuss the most common problem that might occur when we fit a linear regression model to a particular data set.

Resposta: O problema mais comum é o over-fit, ou seja, o nosso modelo está tão bem treinado para o nosso conjunto que, ao tentar prever novos dados ele provavelmente não irá prever direito, pois não possui um comportamento estável.

13. Explain how the simple linear model, given by equation 1, can be improved to yield better prediction accuracy and model interpretability.

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_P\beta_P + \epsilon \tag{10}$$

Resposta: Um jeito de tentar melhorar nossa predição, é trabalhar com modelos penalizados. Como eles serão discutidos mais a frente, eu vou dizer só a ideia geral: Tentamos fazer o modelo não se adequar de maneira excelente ao nosso dataset para poder dar uma 'folga', ou seja, fazer ele ficar mais adaptável e menos enviesado.

14. There are many alternatives to the linear regression model in Equação 10, discuss:

(a) The principal component regression model.

Resposta: O PCA reduz a dimensão dos nossos problemas (excluindo alguns preditores) rotacionando os eixos do problema. Para isso, ele procura os preditores que possuem maior correlação entre os outros para garantir que a perda de informação é mínima. Então, a partir da matriz de correlação, pegamos os autovetores associados aos preditores de maior autovalor e os usamos para rotacionar os eixos.

(b) The partial least squares model.

Resposta: O PLS age praticamente igual ao PCA. A diferença entre eles é que o PLS inclui a saída no modelo. Deste modo, ele calibra entre os preditores que possuem mais informações e que estão mais correlacionados com a saída.

15. For (a) and (b), indicate which of (i.) through (iv.) is correct. Justify your answer.

(a) The lasso, relative to least squares, is:

i. More flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

ii. More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

iii. Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

iv. Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

(b) Repeat (a) for ridge regression relative to least squares.

Resposta: Ambas as respostas são a mesma: **item iii.** Isto ocorre por ambas as penalidades serem bias. Este tipo de modelo de penalidade, no PLS que simplifica o teu modelo ao retirar preditores, tenta aumentar o bias enquanto diminui a variância para evitar que o modelo se enquadre no "over-fit". Um exemplo pode ser visualizado na Figura 2. O modelo em azul seria um modelo em estado 'over-fit', enquanto em vermelho encontraríamos um estado de 'under-fit'. Então os modelos de penalidade tentam equilibrar os

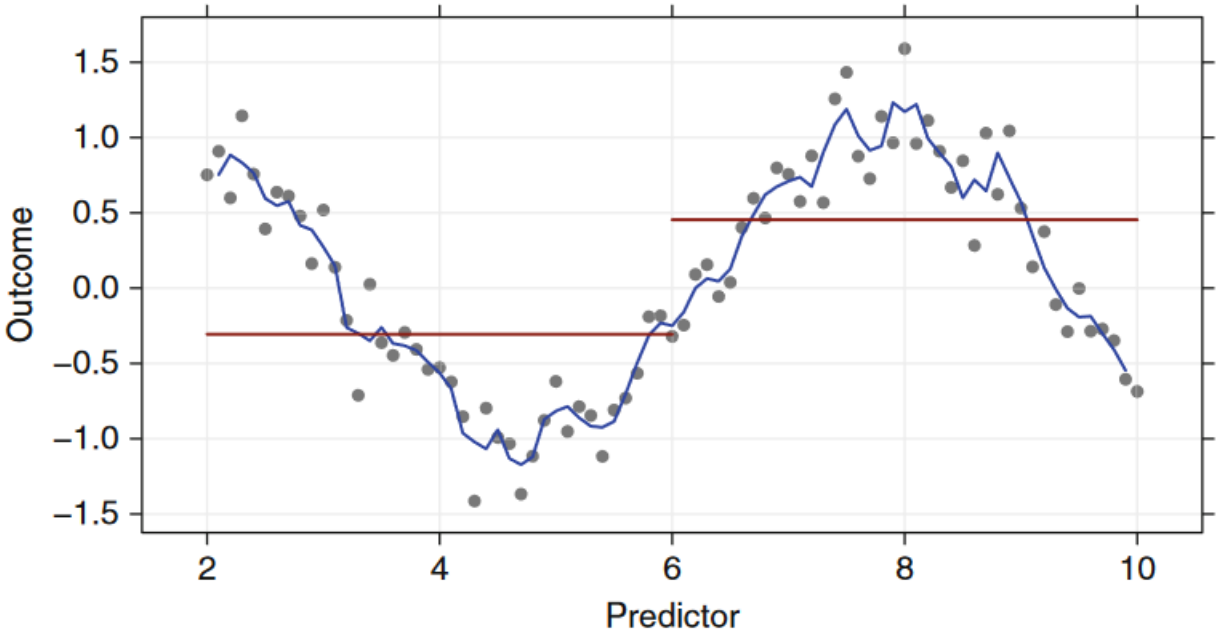

Figura 2: Figura de dois casos extremos. em vermelho um modelo com grande bias e pequena variância, enquanto o modelo em azul possui pequeno bias e alta variância.

dois estados. Também é possível visualizar os estados na Figura 3 e entender mais ou menos a ação desses modelos

16. Suppose that $n = 2, p = 2, x_{11} = x_{12}, x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: $\hat{\beta}_0 = 0$.

(a) Write out the ridge regression optimization problem in this setting.

Resposta: O modelo de penalidade de ridge adiciona o termo no SSE e tenta

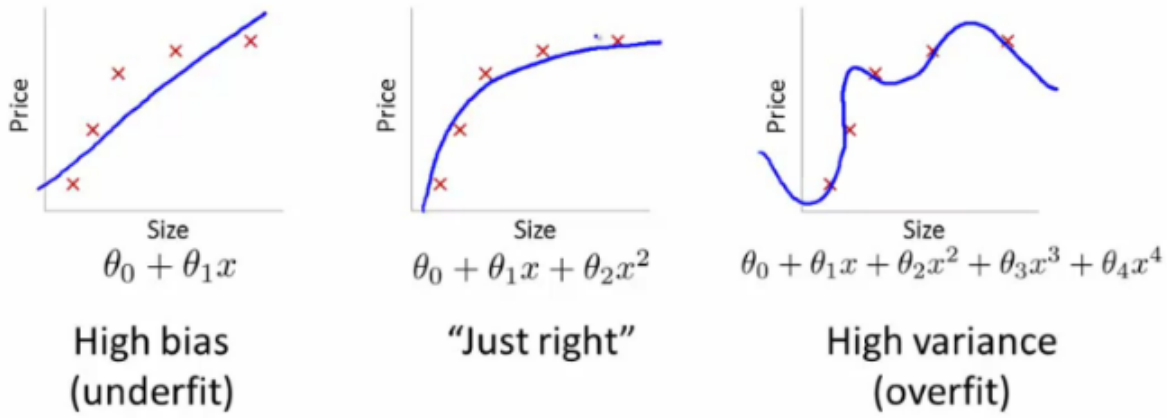

Figura 3: Figura ilustrativa dos estados de underfit, 'ótimo' e overfit

minimizar a seguinte equação:

$$SSE = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) - \lambda\sum_{j=1}^{p}\beta_j^2 \tag{11}$$

Para o caso descrito na questão, conseguimos escrever a seguinte equação:

$$SSE = (y_1 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_1)^2 + (y_2 - \hat{\beta}_1 x_2 - \hat{\beta}_2 x_2)^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2). \tag{12}$$

Onde assumimos que $x_{11} = x_{12} = x_1$ e $x_{21} = x_{22} = x_2$.

(b) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$

Resposta: Fazendo uma derivação parcial da Equação 12 em relação aos betas, achamos o seguinte esquema de equação:

$$\begin{cases} \hat{\beta}_1(x_1^2 + x_2^2 + \lambda) + \hat{\beta}_2(x_1^2 + x_2^2) = y_1x_1 + y_2x_2 \\ \\ \hat{\beta}_1(x_1^2 + x_2^2) + \hat{\beta}_2(x_1^2 + x_2^2 + \lambda) = y_1x_1 + y_2x_2 \end{cases} \tag{13}$$

Subtraindo uma pela outra, é fácil ver que, para $\lambda \neq 0$ a igualdade só ocorre se, e somente se, $\hat{\beta}_1 = \hat{\beta}_2$.

(c) Write out the lasso optimization problem in this setting.

Resposta: Parecido com o ridge, o lasso também adiciona um termo ao SSE e tenta minimizar a seguinte equação:

$$SSE = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) - \lambda\sum_{j=1}^{p}|\beta_j| \tag{14}$$

Para o caso descrito na questão, a equação acima é escrita como:

$$SSE = (y_1 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_1)^2 + (y_2 - \hat{\beta}_1 x_2 - \hat{\beta}_2 x_2)^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|). \tag{15}$$

Onde assumimos que $x_{11} = x_{12} = x_1$ e $x_{21} = x_{22} = x_2$.

(d) Argue that in this setting, the lasso coefficients $\hat{\beta}_1$ e $\hat{\beta}_2$ are not unique - in other words, there are many possible solutions to the optimization problem in (c).

Resposta: Ao colocar as condições dadas no problema, percebemos que $(y_1 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_1)^2 = (y_2 - \hat{\beta}_1 x_2 - \hat{\beta}_2 x_2)^2$. Desta forma, sobra a equação:

$$2[y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_1]^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|) \tag{16}$$

Porém, na penalidade do lasso, nós temos a seguinte condição:

$$|\hat{\beta}_1| + |\hat{\beta}_2| \leq t \tag{17}$$

Desta forma, podemos tratar o problema como minimizar o termo $2[y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_1]^2 \geq 0$. Assim, é fácil que ver que a solução para a equação seria $\hat{\beta}_1 + \hat{\beta}_2 = y_1/x_1$. Porém, temos a condição de contorno que é dada pela Equação 17, assim todos os pontos de interseção da reta $\hat{\beta}_1 + \hat{\beta}_2 = y_1/x_1$ com a condição de contorno (um losango se plotarmos nos eixos β_1 e β_2) são candidatos de parâmetro. Concluímos então que, de fato, os parâmetros não são únicos.

4 Artificial Neural Networks

17. Describe the generic structure of the single perceptron.

Resposta: Um perceptron pode ser descrito como ilustra a Figura 4. Esse modelo simula um neurônio humano: haverá sinais de entrada, os quais serão aplicados pesos W_k em cada input. Também há um peso externo chamado de 'bias' b_k para compor a entrada. Esses dados serão tratados onde é chamado de summing junction, que serve basicamente para fazer a soma ponderada das entradas com seus respectivos pesos. Por fim, essa soma ponderada é mandada para a função de ativação, cujo retorno será nossa própria saída. Deste modo, teremos que nossa saída será dada por: $y = \phi$(Soma Ponderada), onde ϕ é a função de ativação.

18. List and discuss some of the desirable characteristics of artificial neural networks.

19. What is the meaning of learning? Describe the main types of learning strategies.

20. In order to cope with nonlinear problems, additional layer(s) of neurons are placed between the input layer (containing input nodes) and the output neuron are added. Explain the multilayer perceptron architecture.

Resposta: Diferente da Figura 4, haverá outras camadas entre a camada de entrada e a de saída. Estas camadas são chamadas de 'camadas escondidas'. Aqui implementaremos alguns métodos de aprendizado a fim de alterar as funções pesos das entradas para manter a saída com o mínimo de erro possível. Deste

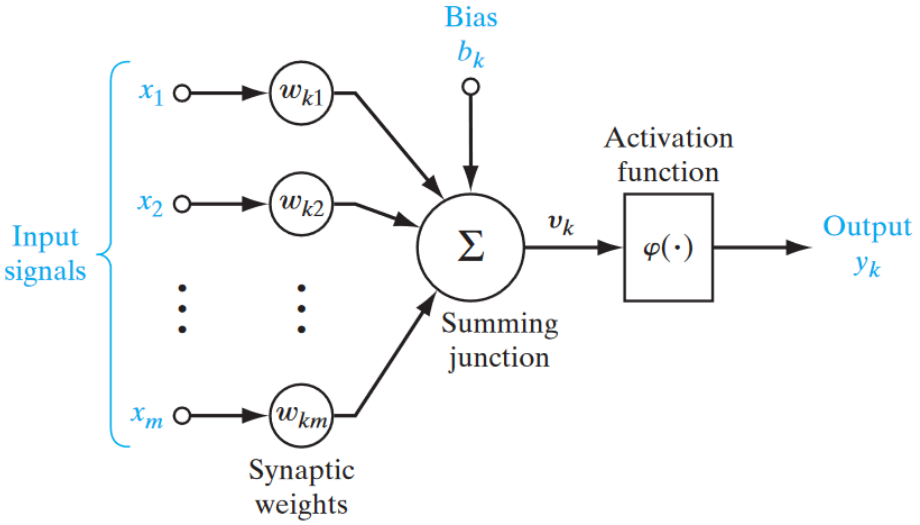

Figura 4: Figura ilustrativa de um perceptron

modo, quanto maior for suas camadas escondidas, mais custoso será para o computador, porém, existirá mais graus de liberdade para tornar o erro mínimo. Na Figura 5 há uma ilustração do que seria uma rede neural de várias camadas.

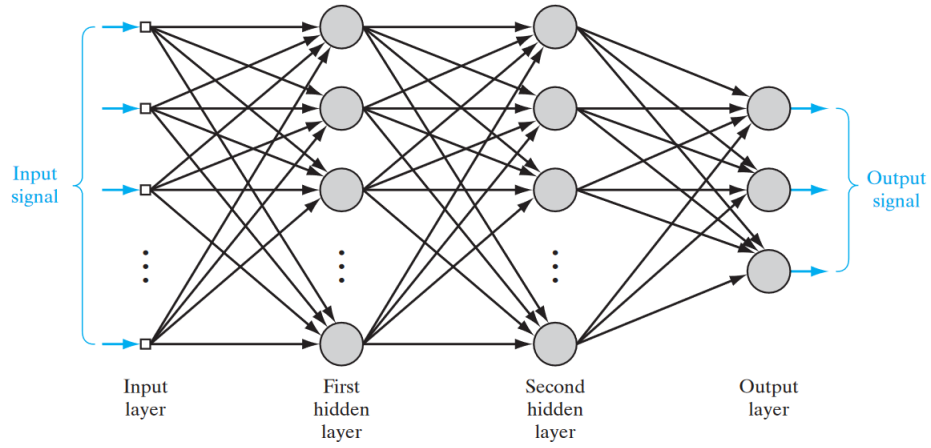

Figura 5: Figura ilustrando uma rede neural de múltiplas camadas

21. Gradient descent optimization has led to one of the most popular learning algoithms, the backpropagation. Derive the basic steps of this algorithm for a single neuron neural network.

22. A sigmoidal neural activation function is standard for feedforward artificial neural networks trained using backpropagation of error. Explain the key advantage of this activation function for such a network in comparison to the alternative linear and step activation functions.

Reposta: Pois durante o algoritmo de backpropagation, precisamos do gradiente. Este, por sua vez, precisa da derivada da função. Deste modo, a função sigmoid

nos dá uma vantagem muito útil: sua derivação é simples. Podemos ver como a função sigmoid (representada por S(x) na Equação 18) e sua derivação se comporta na Equação 18.

$$\begin{aligned} S(x) &= \frac{1}{1+e^{-\lambda x}} \\ \frac{d}{dx}S(x) &= \frac{-1(-\lambda e^{-\lambda x})}{(1+e^{-\lambda x})^2} \\ &= \lambda * \frac{e^{-\lambda x)}}{(1+e^{-\lambda x})} * \frac{1}{(1+e^{-\lambda x})} \\ &= \lambda * (1 - S(x)) * S(x) \end{aligned} \tag{18}$$

5 Classification

23. Why we do not just use least-squares for classification?

Resposta: Ele não é usado pois é muito sensível a outliers. Deste modo, ele não consegue traçar uma linha de divisão adequada. Observe a Figura 6 e perceba que, para a figura a esquerda o LS estava agindo de maneira ótima. Porém, após a adição de outliers (imagem a direita), o LS nos traçou uma reta que não classifica bem nosso modelo.

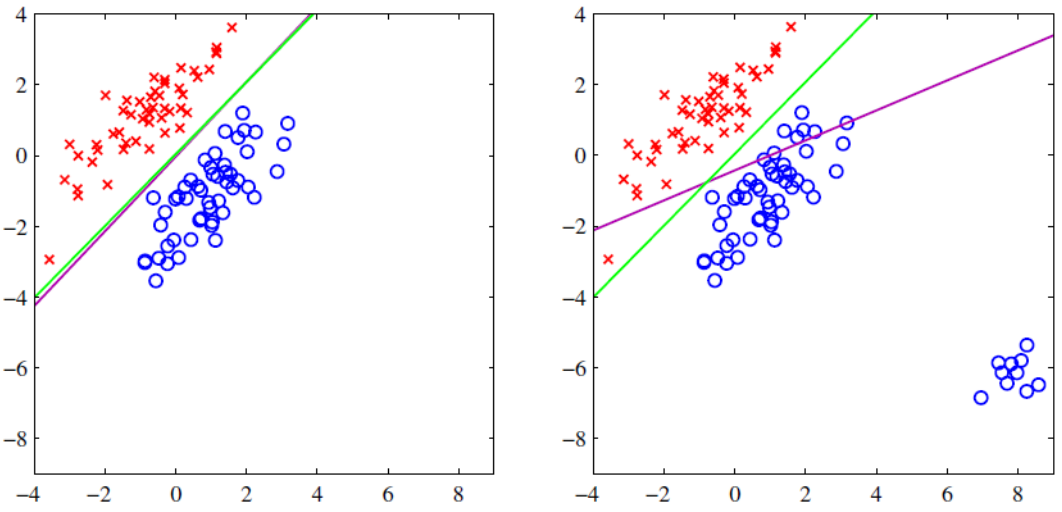

Figura 6: Duas retas de classificação para um determinado dataset. Em verde, temos a reta devido a logistic regression, enquanto em magenta temos a reta devido ao Least-Squares

24. There are many classification techniques that might be used to predict a qualitative response. Among the following carefully describe at least two of them. Highlight their advantages and disadvantages.

    (a) Logistic regression.

    (b) Linear and quadratic discriminant analysis.

    (c) K-nearest neighbors.

    Resposta:

    (a) Logistic Regression: Pode tratar com variações não lineares, visto que pode-se adicionar termos quadráticos, é um modelo de classificação robusto: as variáveis independentes não precisam estar normalizadas, não assume linearidade entre a variável independente e a variável dependente. Entretanto, precisa de muito mais pontos de observação para se estabilizar, também precisamos de um cuidado a mais em identificar as variáveis independentes - já que o modelo dá muita importância a elas -, só consegue classificar problemas binários, ou seja, que há duas possibilidades de saída e, por fim, pode se tornar instável se os dados já tiverem muito bem separados.

    (b) LDA: Ao contrário do Logistic regression, aqui possuímos a possibilidade de fazer predições em saídas de maiores dimensões, precisamos de menos dados para estabilizar e não é instável mesmo se os dados já estiverem bem separados. Por outro lado, precisamos assumir que os dados estão normalizados, ou seja, se distribuem na forma de sino (Bell-Gausian). Além disso, precisamos assumir que todos os inputs de uma mesma classe possuem a mesma média e variância.

    (c) QDA: É a mesma coisa que o LDA, sendo que possui um termo quadrático para lidar com problemas que as classes possuem um alto grau de heterogeneidade.

25. Examine the differences between LDA and QDA.

    (a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

    Resposta: QDA para o data-set de treino, pois é mais flexível que o LDA, assim podemos conseguir melhor precisão. Entretanto, para o de teste, é melhor o LDA devido ao bias-variance tradeoff. Deste modo evitamos overfitting.

    (b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

    Resposta: QDA em ambas as situações, pois possui o termo quadrático para melhor se adaptar a situações não lineares.

(c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

Resposta: Provavelmente aumenta, pois com mais dados o QDA provavelmente irá performar mais para a situação 'ótima' em vez de 'over-fitting'.

26. Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e. K = 1) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

Resposta: Como usamos k=1, no KNN, iremos ter um erro de treino igual a 0%. O enunciado nos diz que a média de erro foi de 18%, concluímos, então, que o erro referente ao caso de teste foi de 36%. Deste modo, é preferível usar Logistic Regression.

6 Genetic algorithms

27. The terms "Genetic algorithms" refer to computer-based problem solving systems that use computational models of evolutionary processes. Explain.

Resposta: Os algorítimos evolucionários são inspirados no processo biológico: seleção natural. Ou seja, lidamos com a adaptação ao meio, mutação, genética, etc. Eles são usados principalmente para resolução de problemas de optimização. É preferido o uso desses algoritmos devido ao fato de que eles usam mais do espaço amostral devido ao seu carácter aleatório - crossover e mutação.

28. Genetic algorithms have a number of components, procedure or operators that must be specified in order to define a particular algorithm. Describe the most important ones.

Resposta:

- Componentes:
  - População: Basicamente o primeiro conjunto de dados e o último. A partir dele que nosso conjunto irá evoluir
  - Parents: Parte selecionada da população que irá gerar os filhos (futura população).
  - Offspring: São os filhos da geração passada, que irão sobreviver ou não e compor a nova população.

- Procedimentos/Operadores:
  - Cromossomo: É onde iremos guardar os genes de cada pessoa no nosso algoritmo. É, geralmente, uma string.
  - Fitness function: Função na qual será usada para selecionar os mais adaptáveis ao meio, para compor a nossa população final ou os selecionados para virarem parents.
  - Seleção: Aqui é o procedimento de quebrar um par de cromossomos (ou seja, escolhe-se dois indivíduos e divide a string cromossomo em duas no mesmo ponto).
  - Cross-Over: Após a quebra do cromossomo na 'seleção', há a troca dos pedaços para gerar dois novos cromossomos.
  - Mutação: Aqui há uma chance pequena de algum valor da nossa string cromossomo ser alterada para um valor aleatório.

29. Provide the general scheme and pseudocode of a genetic algorithm.

    Resposta:

```
INICIA POPULACAO P
ENQUANTO (!CONDICAO DE PARADA):
        SELECAO (P)
        CROSSOVER (P)
        MUTACAO (P)
        FITNESS (P)
FIM_ENQUANTO
```

30. Why should we prefer a population-based heuristic search method (such as a genetic algorithm) to a heuristic search method that lacks a population (such as hill-climbing or simulated annealing)?

    Resposta: Pois o hill-climbing, por exemplo, pode ficar preso em vários pontos: máxima-local, minima-local ou lugares planos como ilustra a Figura 7, enquanto os modelos de population-based não ficam presos, graças ao processo evolucionário.

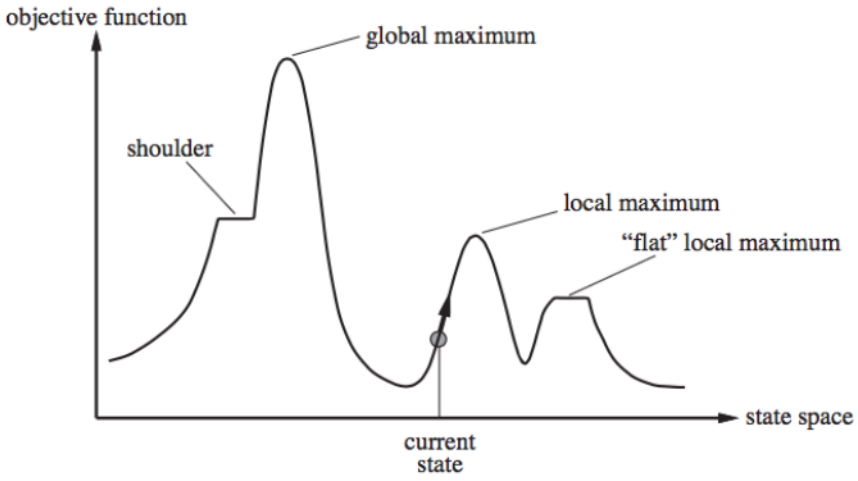

Figura 7: Figura ilustrando pontos onde os algorítimos de hill-climbing fica preso