Chapter 4 – Classification

* No capítulo 3 a gente falou sobre regressão linear, que pode ser utilizada quando temos satisfeitas as condições do erro que falamos no encontro passado
* Mas lá no começo dos nossos estudos falamos sobre as diferenças de entre variáveis quantitativas e qualitativas/categóricas
* Quando falamos de regressão linear, estamos predizendo variáveis quantitativas, para fazermos uma análise de predição de uma varável categórica, dizemos que temos um problema de classificação (exemplo da cor do cabelo)
* O capítulo vai entrar na discussão de 3 dos classificadores categóricos mais utilizados:
  + Regressão logística
  + LDA – Linear discriminant analysis
  + KNN – o famoso K-nearest neighbors, abordado no cap 2
* Podemos falar do exemplo dos classificadores da página 129 (bom pra fixar a ideia)
* Citar o conceito de dummy variable -> Quando “forçamos” um problema de classificação a caber em uma análise quantitativa. Já foi visto no cap 3
* Regressão logística
  + Quando a gente fala de uma regressão logística, estamos falando da probabilidade da variável sob análise (no caso do livro, Y) estar em uma das classificações, ao invés de tentar predizer o valor da variável
  + Como estamos falando da probabilidade, faz sentido que utilizemos uma modelagem que considere os valores entre 0 e 1 -> Usaremos o modelo logístico (tem várias funções que poderiam satisfazer essa simples condição)
  + Mostrar um pouco a manipulação até o ponto em que tiramos uma relação linear com o logit. A partir daí, iremos encontrar Beta0 e Beta1, que são os estimadores -> através dos dados de treino,ào claro
  + Fala um pouco sobre a manipulação da função de probabilidade (parte dos produtórios)
  + O Z estatístico aí então vai cumprir o mesmo papel que o t estatístico cumpria lá na regressão linear. Por exemplo, se tiver um z muito grande, é um indicativo de pouca ou nenhuma dependência
  + Uma vez definido os estimadores, Beta0 e Beta1, será possível fazer as predições
  + Para múltiplos preditores, a manipulação será a mesma, importante analisar diversos preditores e seu z estatístico antes de sair fazer uma regressão logística com um preditor que você acha apropriado
  + Quando tratamos de regressão logística, geralmente estamos lidando com problemas de 2 classes (ser loiro ou não loiro, por exemplo). Quando lidamos com mais classes, iremos lançar mão de outros classificadores, como o LDA
* LDA – Linear Discriminant Analysis
  + Iremos utilizar quando as classes forem bem separadas (ex: ser loiro, moreno, ruivo) – Nesse caso os estimadores são muito instáveis em uma regressão logística
  + Aqui vai entrar o conceito do teorema de Bayes, que também foi visto no cap 2, aliás, já foi visto lá que o classificador de Bayes vai, tendo definida a probabilidade da observação estar em cada uma das classes, vai aloca-la para a classe com maior probabilidade
  + 4.4.2 vai analisar LDA quando p=1, ou seja, só temos um preditor. E tentar encontrar uma estimativa para a função densidade do x pra classe k. Aqui é importante essa hipótese quando a gente lida com LDA, que está logo no começo do capítulo, essa função densidade fk(x) é normal (ou próxima de uma distribuição normal). -> take a look na eq 4.11
  + Também é assumido que a variância é compartilhada por todas as classes.
  + Na figura 4.4. ele vai mostrar o conceito de se traçar a “fronteira de decisão de Bayes” -> importante comentar de como as hipóteses estão sendo levadas em consideração nesse exemplo para a definição da fronteira
  + Agora falando do LDA, a gente vai pegar o teorema de Bayes e aplicar as estimativas das eqs de 4.15
  + Essa função aí com tudo substituído é a função de discriminação, que vai classificar a observação na classe que maximizá-la
  + 4.4.3 agora vai expandir esse conceito que acabamos de falar para quando temos p>1, ou seja, múltiplos preditores, mas ainda temos a hipóteses das distribuições serem normais
  + Daí, encontramos na eq 4.19 a expansão do que vimos para p=1 na eq. 4.17
  + Consideração importante aqui, começa a ficar bem claro, até pelas figuras, que quanto maior a razão parâmetros/amostras mais nítido será o fenômeno de overfitting
  + Ali na tabela 4.4 irá entrar o conceito de matriz de confusão, que utilizamos até em outros contextos -> Falso, Verdadeiro x Positivo, Negativo
  + Depois, toma-se algumas discussões do exemplo do default que foi utilizado, muito específicas para uma explicação geral
  + Na QDA, Quadratic Discriminant Analysis, as hipóteses são as mesmas EXCETO que na QDA cada classe tem sua matriz de covariância, então aquela equação que maximizamos (4.17, 4.19) agora vai ficar no formato da 4.23
  + Por que iremos utilizar uma abordagem de que cada classe tem sua matriz de covariância? -> Uma questão do trade-off que sempre falamos de viés-variância, principalmente quando começamos a estudar overfitting.
  + Em resumo, LDA será mais preciso, mas terá muito viés para casos em que o número de amostras for muito grande. QDA será menos preciso, mas será melhor com o maior número de amostras. Olhar a explicação gráfica disso na figura 4.9.
* 4.5 é uma comparação final dos métodos abordados no capítulo.
* Esse capítulo não fala do KNN novamente, mas é importante lembrar que ele se comportava muito bem quando nós tínhamos fronteiras de classes “muito não lineares” MAS é uma abordagem não paramétrica ==== Não sabemos quais preditores são mais relevantes (só pensar em quais situações isso seria importante ou não).
* Importante então colocar numa escala de flexibilidade/viés 🡺 Regressão Logística/LDA – QDA – KNN, vai depender muito da natureza da classificação, do dataset que você tem disponível e pode ser até ideal tentar usar os diferentes classificadores e ver qual melhor se adapta ao problema que você está tentando resolver.