

**Aprendizado Supervisionado Modelos Lineares**

**Prof. Raphael Carvalho**

**Introdução**

● Uma das principais aplicações de aprendizado de máquina é a previsão, isto é, quando queremos prever algum atributo tendo somente alguns dados de

entrada.

● Determinamos como fazer essa previsão com base em exemplos históricos de dados de entrada e saída, isto é, baseado em comportamentos observados no passado, conseguimos fazer inferências sobre o futuro.

**2**

**Modelo Preditivo**

● É aquele responsável por relacionar dados de entrada (variáveis independentes) com o resultado esperado (variável dependente ou variável alvo contínua)

● Diferentes modelos geram formas matematicamente muito diferentes de construir a relação entre as variáveis de entrada e de saída, tornando-os assim capazes de captar padrões estatísticos também

diferentes.

**3**

**Modelo Preditivo**

● Em regra, é preciso realizar experimentos computacionais, avaliando o desempenho de modelos de tipos diferentes para descobrir qual é o mais adequado a uma tarefa e um conjunto específico de dados. Por que?

● Porque cada tipo de modelo tem suas características, seus pontos fortes e fracos e sua lógica de funcionamento. ● Não é preciso reimplementar um algoritmo do zero para entender suas propriedades fundamentais e utilizá-lo adequadamente

**4**

**Modelos Preditivos**

● São basicamente uma função matemática que, quando aplicada a uma massa de dados, é capaz de identificar padrões e oferecer uma previsão do que pode ocorrer.

● Existem vários tipos de modelos de predição, dentre eles se destacam os modelos lineares.

● Esses modelos compreendem uma ampla família de modelos derivados da estatística, embora apenas dois deles (regressão linear e a regressão logística) sejam frequentemente utilizados.

**5**

**Modelos Lineares**

● São fáceis de entender, rápidos de criar e moleza de implementar do zero.

● Se você os dominar, você realmente tem o equivalente a um canivete suíço para aprendizado de máquina que não pode fazer tudo perfeitamente, mas pode

atendê-lo prontamente e com excelentes resultados. ● Os dois modelos lineares mais conhecidos são a regressão linear e a regressão logística.

**6**

**Regressão Linear**

7

**Regressão Linear**

● É a ferramenta estatística que nos ajuda a quantificar a relação entre uma variável específica e um resultado que nos interessa enquanto controlamos outros fatores.

● Podemos isolar o efeito de uma variável enquanto mantemos os efeitos das outras variáveis constantes. ● A imensa maioria dos estudos que você lê nos jornais é baseada em análise de regressão.

**8**

**Regressão Linear**

● Em essência, a regressão linear busca encontrar o “melhor encaixe” para uma relação linear entre duas variáveis.

**9**

**Exemplo de Regressão Linear**

● Relação entre altura e peso.



**10**

**Exemplo de Regressão Linear**

● Se lhe fosse pedido que descrevesse o padrão, você poderia dizer algo mais ou menos do tipo: “O peso parece aumentar com a altura”.

● A regressão linear nos dá a possibilidade de ir além e “encaixar uma reta” que melhor descreva uma relação linear entre as duas variáveis (Peso e Altura)

● Muitas retas possíveis são amplamente consistentes com os dados de altura e peso, mas como sabemos qual é a melhor reta para esses dados?

**11**

**Exemplo de Regressão Linear**

● Possível resultado de uma reta



**12**

**Como resolver?**

● É aqui que entra em cena o aprendizado de máquina! ● A ideia do algoritmo é oferecer vários dados para que ele encontre a equação que melhor descreve e se ajusta aos dados, isto é, que minimize a variância dos erros em uma predição.

● A Regressão Linear utiliza tipicamente uma

metodologia chamada de Mínimos Quadrados

Ordinários (MQO)

**13**

**Como resolver?**

● MQO encaixa a reta que minimiza a soma dos residuais elevados ao quadrado.

● Residual é a distância vertical a partir da reta de regressão, exceto para aquelas observações que se situam diretamente em cima da reta, para as quais o residual vale zero.

● A fórmula pega o quadrado de cada residual antes de somar todos e isso aumenta o peso dado àquelas observações mais distantes da reta de regressão – chamadas de extremos ou outliers.

● Dessa forma, os mínimos quadrados ordinários “encaixam” a reta que minimiza a soma dos residuais ao quadrado conforme é ilustrado na imagem da página anterior.

**14**

**Resultados**

● Os mínimos quadrados ordinários nos dão a melhor descrição de uma relação linear entre duas variáveis. ● O resultado não é somente uma reta, mas uma equação que descreve essa reta.

● Essa equação é conhecida como equação de regressão linear simples e assume a seguinte forma:

y = a + bx ou y = α + ßx

**15**

**Equação de regressão linear simples**

y = a + bx

● **y**: peso em quilos;

● **a**: intercepto, isto é, ponto em que a reta intercepta o eixo y (valor de y quando x = 0)

● **b**: inclinação da reta

● **x**: altura em centímetros.

**16**

**Analisando o resultado**

● A inclinação da reta que encaixamos descreve a “melhor” relação linear entre altura e peso para essa amostra, conforme definida pelos mínimos quadrados ordinários. ● A reta de regressão é perfeita?

● É claro que não!

● Ela com certeza não descreve perfeitamente toda observação nos dados.

**17**

**Regressão Logística**

18

**Regressão Logística**

● É um algoritmo de aprendizagem de máquina supervisionado utilizado para classificação

● Em geral, a utilização da regressão logística se dá com categorias binárias, isto é, aquelas que podem assumir somente dois valores

● Ex: grande ou pequeno, alto ou baixo, sim ou não, lucro ou prejuízo, válido ou inválido...

**19**

**Regressão Logística**

● Vamos imaginar que queiramos definir se um determinado paciente está ou não infectado com coronavírus

● Para tal, vamos reunir diversas informações contidas em um exame de sangue como contagem de anticorpos, contagem de plaquetas, contagem de leucócitos (variáveis independentes)

● Em seguida, aplica-se um coeficiente ou peso a cada uma dessas variáveis que comporão uma função de regressão linear múltipla e retornará um determinado valor como resposta (variável dependente).

**20**

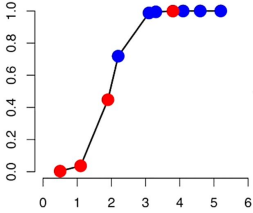
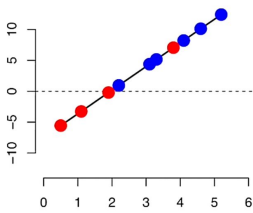
**Regressão Logística**

● Ocorre que a regressão logística é um tipo de algoritmo de classificação, logo precisamos transformar esse valor real retornado pela regressão linear em uma das categorias pré-definidas por um supervisor.

● Para tal, temos que utilizar um modelo logístico para fazer um mapeamento desse valor dentro de um intervalo entre [0, 1], que pode ser interpretado como a probabilidade de ser da categoria que nos interessa. **21**

**Função Sigmóide**

● Função de ativação que recebe como entrada um número real [-∞, +∞] e retorna um número entre [0,1]

**22**

**Como isso é feito?**

● Com o conjunto de casos em que sabemos se a pessoa estava infectada ou não, podemos treinar o modelo de modo que possa ir ajustando até encontrar um valor razoável.

● O modelo calcula os coeficientes das variáveis

independentes da minha regressão linear para refletir um valor coerente de probabilidade após aplicar a regressão logística

**23**

**Como isso é feito?**

● Em seguida, coloca-se os valores de quantidade de anticorpos, leucócitos e plaquetas de uma pessoa que está com coronavírus, o modelo ajusta os coeficientes (a, b, c, d) e retorna um resultado.

● Depois, após aplicar a função sigmóide nesse resultado, espera-se que retorne uma alta probabilidade de essa pessoa estar com coronavírus (Ex: 0,9), dado que a pessoa realmente está infectada com coronavírus, logo esse seria um resultado coerente.

● Se retornar um valor como 0,2 (isto é, 20% de probabilidade de essa pessoa estar infectada com coronavírus), significa que o modelo ainda não está bom pois sabemos que pessoa efetivamente está infectada com coronavírus

**24**