Aprendizagem de Máquina

Métodos Generativos

Paulo Manoel da Silva Junior
paulomanoel14@gmail.com
www.de.ufpb.br

Sumário

Introdução

Observação

Esse resumo foi realizado através do livro **An Introduction to Statistical Learning: with Applications in R. 2nd ed., Springer, 2021.**

- A regressão logística envolve modelar diretamente $P(Y = k|X = x)$, levando em consideração apenas o caso de duas respostas, ou uma variável categórica com dois níveis.

Introdução

- Agora, vamos considerar uma abordagem alternativa e menos direta para estimar essas probabilidades. Nesse novo enfoque, vamos modelar a distribuição das variáveis preditoras, *ou features*, separadamente em cada uma das classes de resposta (Ou seja, para cada valor de Y), que é nossa variável *target*. Em seguida, utilizamos o teorema de Bayes para reverter essa informação e obter estimativas para $P(Y = k|X = x)$. Quando a distribuição de X dentro de cada classe é assumida como normal, descobre-se que o modelo é muito semelhante em forma à regressão logística

Introdução

- Por que precisamos de outro método, se já temos a regressão logística?
  Alguns motivos
  1. Quando há uma separação substancial entre as duas classes, as estimativas dos parâmetros para o modelo de regressão logística são surpreendentemente instáveis. Os métodos que consideramos nesta seção não sofrem desse problema.
  2. Os métodos desta seção podem ser naturalmente estendidos para o caso de mais do que duas classes de resposta. (No caso de mais do que duas classes, a regressão logística pode ser generalizada usando técnicas como a regressão logística multinomial, mas outros métodos podem oferecer vantagens específicas em determinados cenários.)

Análise Discriminante Linear para $p = 1$

Vamos supor no momento que $p = 1$, ou seja, que temos apenas uma variável preditora. E o que gostaríamos de obter é uma estimativa para $f_k(x)$ que possamos inserir no teorema de Bayes, *que encontra-se abaixo* para estimar $P_k(x)$. Em seguida, classificaremos uma observação na classe para a qual $P_k(x)$ é maior. Para estimar $f_k(x)$, precisamos de algumas suposições sobre a sua forma.

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{1}$$

Supondo normalidade no caso unidimensional. A densidade da normal fica da seguinte forma.

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right) \tag{2}$$

Análise Discriminante Linear para $p = 1$

Por enquanto, vamos assumir que $\sigma_1^2 = \cdots = \sigma_k^2$, ou seja, há um termo de variância compartilhado entre todas as K classes, que podemos denotar apenas de σ^2. Então, substituindo a equação da normal acima, no teorema de Bayes, ficamos com a seguinte expressão.

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)} \tag{3}$$

Observe que na equação 3, π_k denota a probabilidade a priori de que uma observação pertença à k-ésima classe. Ao fazer algumas transformações e aplicar o logaritmo, podemos verificar que é o mesmo que atribuir uma observação a classe equivale a:

Análise Discriminante Linear para $p = 1$

$$\delta_k = x \times \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k) \tag{4}$$

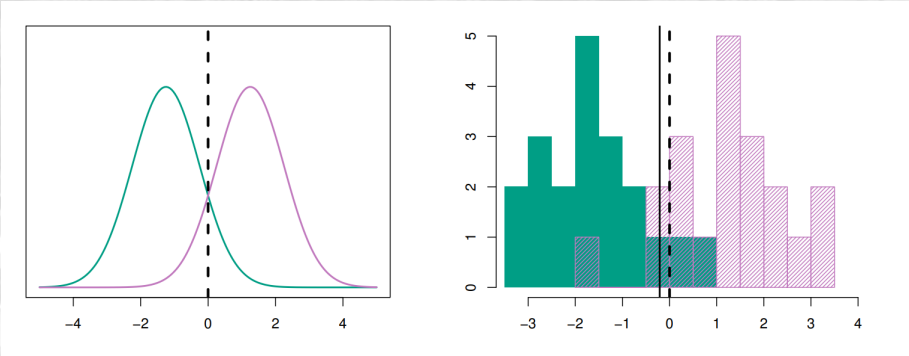
- Meramente, consiste em atribuir a classe a qual essa quantidade acima é maior. Por exemplo, se $K = 2$ e $\pi_1 = \pi_2$, então o classificador de Bayes, atribui a observação a classe 1 se $2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$, e a classe 2 caso contrário. A fronteira de decisão de Bayes é o ponto para o qual $\delta_1(x) = \delta_2(x)$, e pode se mostrar que isso equivale a:

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2} \tag{5}$$

- Isso pode ser ilustrado com a figura a seguir.

Análise Discriminante Linear para $p = 1$

Figura: Densidades da normal e acréscimo da fronteira de decisão com utilização de LDA



Fonte: Retirada do Livro.

Análise Discriminante Linear para $p = 1$

- Na prática, mesmo sabendo que nossa suposição de que X é retirado de uma distribuição gaussiana dentro de cada classe, mesmo assim para aplicar o classificador de Bayes, precisamos estimar alguns parâmetros $\mu_1, \cdots, \mu_K, \pi_1, \cdots, \pi_K$ e σ^2. O método de **Análise de Discriminante Linear**, aproxima o classificador de Bayes ao inserir estimativas para π_k, μ_k, e σ^2. Em particular, são utilizados as seguintes estimativas:

$$\hat{\mu_k} = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{6}$$

$$\hat{\sigma^2} = \frac{1}{n-K} \sum_{K=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu_k})^2 \tag{7}$$

Análise Discriminante Linea para $p = 1$

- Onde n é o número total de observações no conjunto de treinamento, n_k é o número total de observações no conjunto de treinamento de determinada classe. A estimativa para μ_k é simplesmente a média de todas as observações de um treinamento da k-ésima classe, enquanto o σ^2, pode ser visto como uma média ponderada das variâncias amostrais para cada uma das K classes. Às vezes o conhecimento a priori das probabilidades de pertencimento a cada classe pode ser utilizado. Na ausência dessas informações, o método estima π_k usando a proporção das observações de treinamento. Ficando dessa maneira.

$$\hat{\pi_k} = \frac{n_k}{n} \tag{8}$$

Análise Discriminante Linear para $p = 1$

Então, o classificador insere as estimativas fornecidas na Equação 6, 7 e 8 na equação 4 e atribui uma observação X = x à classe para a qual

$$\hat{\delta}_k(x) = x \times \frac{\hat{\mu_k}}{\hat{\sigma}^2} - \frac{\hat{\mu_k}^2}{2\hat{\sigma}^2} + log(\hat{\pi_k}) \tag{9}$$

Seja o maior valor r.
A palavra linear do classificador, deriva do fato de que as funções discriminantes são lineares, ou seja, a fronteira de decisão será linear, como no caso observado na Figura, onde a fronteira de decisão era uma reta.

Análise Discriminante Linear para $p > 1$

- Mas, e se tivermos o caso de mais de uma variável preditora?
- Solução: Podemos estender o classificador, para isso assumimos que $X = (X_1, X_2, \cdots, X$ sendo retirado por exemplo de uma normal multivariada e específico de uma matriz de covariância comum.
- O autor faz uma revisão da normal multivariada (*Assunto já visto em outras disciplinas*)

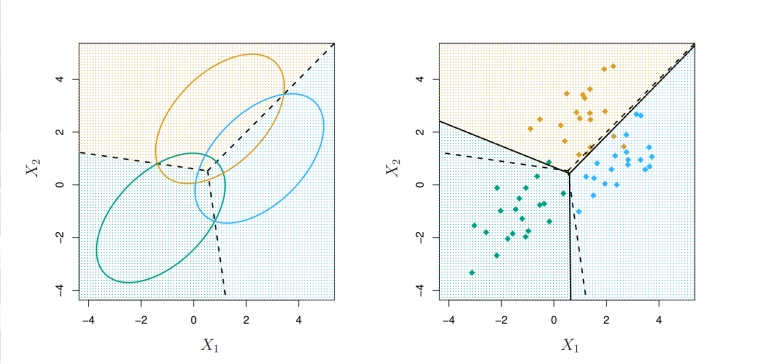
Análise Discriminante Linear para $p > 1$

- No caso de $p > 1$ preditores, o classificador assume que as observações na k-ésima classe são retiradas de uma distribuição normal multivariada $N(\mu_k, \Sigma)$, onde μ_k é um vetor de média da classe específica, e Σ é uma matriz de covariância comum a todas as K classes. Ao inserir a função de densidade para a k-ésima classe, $f_k(X = x)$, na equação 1 (Teorema de Bayes) e realizando um pouco de álgebra, revela-se que o classificador de Bayes atribui uma observação $X = x$ à classe para a qual essa quantidade da equação abaixo é maior.

$$\Gamma_k = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log \pi_k \tag{10}$$

Análise Discriminante Linear para $p > 1$

Figura: Ilustração de uma normal multivariada com duas variáveis preditoras, linha tracejada classificador de Bayes, linha sólida classificador LDA



Fonte: Retirada do Livro.

Análise Discriminante Linear para $p > 1$

- O autor traz o exemplo de uma empresa que concede crédito e quer classificar seus possíveis clientes em inadimplentes ou adimplentes. Com duas variáveis preditoras, e mostra que o classificador chega bem próximo ao classificador de Bayes, pois, o classificador de Bayes, consiste em atribuir uma informação a classe, onde a probabilidade daquela informação específica pertencer a determinada classe seja maior do que 0.5.

Análise Discriminante Linear para $p > 1$

- Só que ele traz possíveis problemas neste exemplo
  1. Os dados foram bem ajustados para o conjunto de teste, trazer uma margem de erro geral de 3.73%, isso pode causar *overfitting*.
  2. Quando utilizado o valor de 0.5 como o valor delimitador para classificar um cliente como inadimplente, o percentual de verdadeiros inadimplentes que o método conseguia classificar corretamente era de 24.3%. Então, como possível solução ele trouxe a redefinição do valor limite para classificar um cliente em inadimplente, que seria 0.20, ou seja, se a probabilidade de que ele fosse inadimplente for maior ou igual a 0.20, ele deve ser categorizado em tal classe.

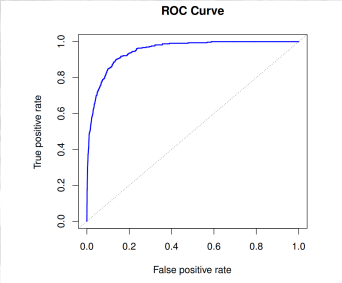
Análise Discriminante Linear para $p > 1$

- No entanto, quando o valor de decisão é alterado a taxa de erro geral que era de 3.73% tem um ligeiro acréscimo, todavia, a taxa de acertos para classificar como inadimplentes, aumenta.
- E agora, o que podemos utilizar para decidir qual seria o melhor valor para considerar como o limiar, a curva ROC (Curva característica de operação) pode ser uma boa aliada.
- Na curva ROC, temos:
  - **A taxa de verdadeiros positivos:** Conhecida como sensibilidade, que é a taxa de inadimplentes corretamente identificados
  - **A taxa de falsos positivos:** Conhecida como especificidade, que é a fração de não inadimplentes que foi classificada incorretamente como inadimplentes.

Análise Discriminante Linear para $p > 1$

- A curva ROC ideal abraça o canto superior esquerdo, indicando uma alta taxa de verdadeiros positivos e uma baixa taxa de falsos positivos.

Figura: Exemplo da Curva ROC



Fonte: Retirada do Livro.

Análise Discriminante Linear para $p > 1$

- O desempenho geral de um classificador, resumido por todos os limiares possíveis, é dado pela área sob a curva ROC (AUC). Uma curva ROC ideal ficará próxima do canto superior esquerdo, assim, quanto maior a área sob a curva ROC (AUC), melhor o classificador. Para esses dados, o AUC é de 0,95, o que é próximo do máximo de 1,0, sendo considerado muito bom. Esperamos que um classificador que não performa melhor do que o acaso tenha um AUC de 0,5 (quando avaliado em um conjunto de testes independente não utilizado no treinamento do modelo).
- Curvas ROC são úteis para comparar diferentes classificadores, já que levam em conta todos os limiares possíveis.

Análise Discriminante Quadrática

- Assim como a análise discriminante linear, a QDA resulta da suposição de que as observações são retiradas de uma distribuição gaussiana e da aplicação das estimativas dos parâmetros ao teorema de Bayes para realizar as previsões. No entanto, o QDA assume que cada classe possui a sua própria matriz de covariância. Ou seja, pressupõe que uma observação da k-ésima classe segue $X \sim N(\mu_k, \Sigma_k)$, onde Σ_k é uma matriz de covariância para a k-ésima classe. Sob essa suposição, o classificado Bayesiano atribui uma observação $X = x$, à classe para a qual

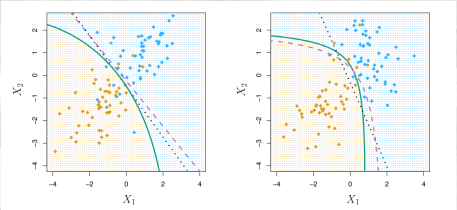
$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}log \mid \Sigma_k \mid + log\pi_k$$
$$= -\frac{1}{2}x^T\sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}log \mid \Sigma_k \mid + log\pi_k \quad (11)$$

Análise Discriminante Quadrática

- Bem, então consiste a atribuir a observação a classe a qual o valor da Equação 11 é o maior.
- Porque importa se assumimos ou não que as K classes compartilham uma matriz de covariância comum? Em outras palavras, por que alguém preferiria LDA em relação ao QDA, ou vice-versa? A resposta está no trade-off entre viés e variância. Quando há p preditores, estimar uma matriz de covariância requer a estimativa de $p(p+1)/2$ parâmetros. O QDA estima uma matriz de covariância separada para cada classe, totalizando $Kp(p+1)/2$ parâmetros. Com 50 preditores, isso é um múltiplo de 1.275, o que representa muitos parâmetros.
- O LDA é um classificador menos flexível do que o QDA e, portanto, tem uma variância substancialmente menor.

Análise Discriminante Quadrática

- O QDA é recomendado se o conjunto de treinamento for muito grande, de modo que a variância do classificador não seja uma preocupação principal, ou se a suposição de uma única matriz de covariância comum para as K classes for insustentável.
- A Figura abaixo ilustra o desempenho do LDA e do QDA em dois cenários, sendo assim com a figura abaixo podemos verificar que como a fronteira de decisão agora é quadrática, o QDA tem um desempenho melhor do que o classificador linear.



Fonte: Retirada do Livro.

Naive Bayes

- O classificador Naive Bayes adota uma abordagem diferente para estimar a função $f_1(x), \cdots, f_K(x)$. Em vez de assumir que essas funções pertencem a uma família específica de distribuições, é feito uma única suposição: **Dentro da k-ésima classe, os p preditores são independentes.**
- Matematicamente pode ser representada da seguinte forma

$$f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p) \tag{12}$$

- Onde f_{kj} é a densidade do j-ésimo preditor entre observações na k-ésima classe.

Naive Bayes

- Por que essa suposição é tão poderosa? Essencialmente, estimar uma função de densidade p-dimensional é desafiador porque precisamos considerar não apenas a distribuição marginal de cada preditor - isto é, a distribuição de cada preditor isoladamente - mas também a distribuição conjunta dos preditores. No caso de uma distribuição normal multivariada, a associação entre os diferentes preditores é resumida pelos elementos fora da diagonal da matriz de covariância. No entanto, em geral, essa associação ser muito dificil de caracterizar e extremamente desafiadora de estimar.

Naive Bayes

- Características do Naive Bayes
  1. Nem sempre e na maioria dos casos a suposição de que as p covariáveis são independentes é sustentada, todavia, mesmo assim em alguns casos o método leva a resultados bastante decentes, especialmente em configurações onde n não é grande o suficiente em relação a p para que possamos estimar efetivamente a distribuição conjunta dos preditores dentro de cada classe.
  2. Essencialmente. a suposição do Naive Bayes introduz algum viés, mas reduz a variância, resultando em um classificador que funciona bastante bem na prática como resultado do trade-off entre viés e variância.

Naive Bayes

- Assim que fazemos a suposição de Naive Bayes, podemos substituir a equação 12 na equação 1 para obter uma expressão para a probabilidade posterior

$$P(Y = k|X = x) = \frac{\pi_k \times f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_k \times f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)} \quad (13)$$

- para $k = 1, \cdots, K$
- Se X_j for quantitativo, podemos assumir que $X_j|Y = k \sim N(\mu_{jk}, \sigma_{jk}^2)$. Em outras palavras, assumimos que dentro de cada classe, o preditor j é extraído de uma normal (univariada). Embora isso possa soar um pouco como o QDA, há uma diferença fundamental, pois aqui estamos pressupondo que os preditores são independentes; isso equivale ao QDA com uma suposição adicional de que a matriz de covariância específica da classe é diagonal.

Naive Bayes

- Se X_J for quantitativo, outra opção é usar uma estimativa não paramátrica para f_{kj}. Uma maneira muito simples de fazer isso é criar um histograma para as observações do j-ésimo preditor dentro de cada classe. Então podemos estimar $f_{kj}(x_j)$ como a fração das observações de treinamento na k-ésima classe que pertencem ao mesmo intervalo de histograma que x_j. Alternativamente, podemos usar um estimador de densidade da kernel, que é essencialmente uma versão suavizada de um histograma.

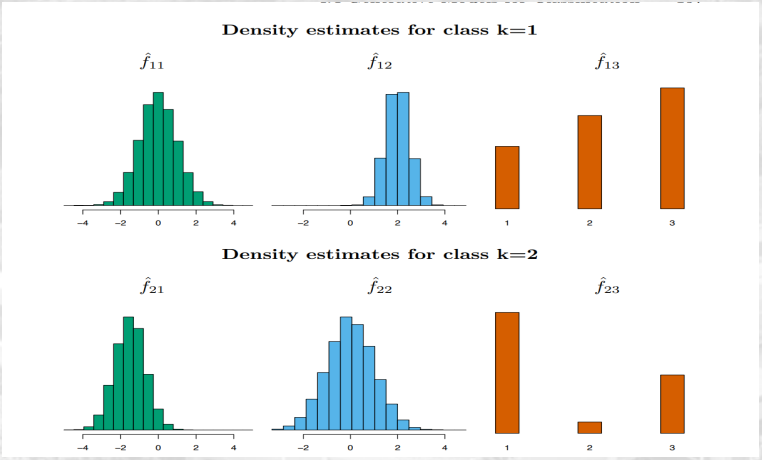
Naive Bayes

- Se X_j for qualitativo, podemos simplesmente contar a proporção de observações de treinamento para o j-ésimo preditor correspondente a cada classe. Por exemplo, suponha que $X_j \in \{1, 2, 3\}$ e temos 100 observações na k-ésima classe. Suponha que o j-ésimo preditor assuma valores de 1,2, e 3 em 32, 55 3 13 dessas observações, respectivamente. Então, podemos estimar f_{kj} como:

$$\hat{f}_{kj}(x_j) = \begin{cases} 0.32 & se\ x_j = 1 \\ 0.55 & se\ x_j = 2 \\ 0.13 & se\ x_j = 3 \end{cases} \tag{14}$$

Naive Bayes

Figura: Ilustração



Fonte: Retirada do Livro

Aprendizagem de Máquina

Métodos Generativos

Paulo Manoel da Silva Junior
paulomanoel14@gmail.com
www.de.ufpb.br