CAPÍTULO 2 - APRENDIZADO ESTATÍSTICO

2.1. O que é aprendizado estatístico?

Para ilustrarmos nosso estudo de aprendizado estatístico, comecemos com um exemplo simples. Suponha que somos consultores de estatística contratados por um clientepara prover consultoria em como aumentar as vendas de um produto em particular. O conjunto de dados $Propagandas$ consiste das vendas daquele produto em 200 mercados diferentes,juntamente com orçamentos de propagandas para cada um desses mercados para três diferentes mídias: TV, $rádio$ e $jornal$. Não é possível para nosso cliente aumentar diretamente as vendas do produto. Por outro lado, é possível controlar os gastos em cada uma das três mídias. Portanto, se determinarmos que há uma associação entre propagandas e vendas, podemos instruir nosso cliente a ajustar os orçamentos de propagandas, aumentando então as vendas. Em outras palavras, nosso objetivo é desenvolver um modelo preciso que possa ser utilizado para prever vendas com base nos três orçamentos de mídia. Neste contexto, os orçamentos de propagandas são variáveis de entrada enquanto $vendas$ são variáveis de saída. As variáveis de entrada são tipicamente denotadas por X. X_1 seria o orçamento da TV, X_2 o orçamento do $rádio$, e X_3 o orçamento do $jornal$. As entradas são chamadas por diversos nomes, como variáveis indepenndentes, preditores, características, ou apenas variáveis. A variável de saída (no caso "vendas) é frequentemente chamdada de resposta ou variável dependente, e é tipicamente denotada por Y.

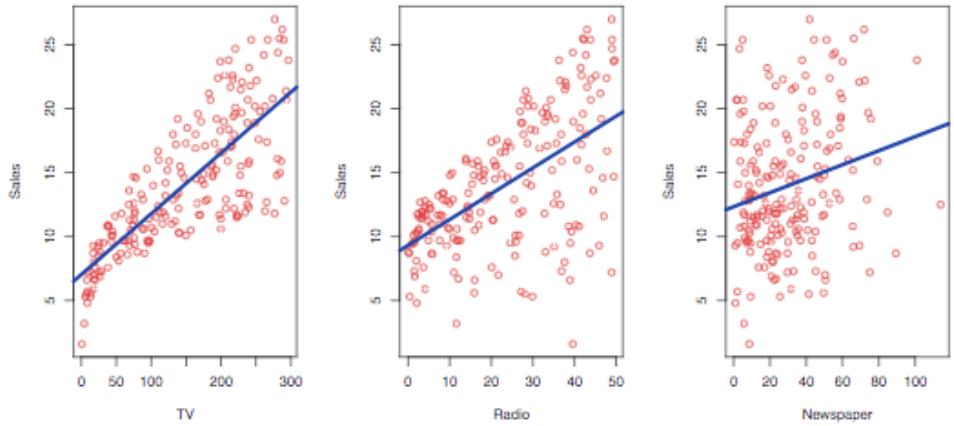

Figura 2.1.
O dataset $Propagandas$. O gráfico mostra $vendas$, em milhares de unidades, como função do orçamento para TV, $rádio$ e $jornal$, em milhares de dólares, para 200 mercados diferentes. Em cada gráfico mostramos o ajuste por mínimos quadrados simples de $vendas$ para cada variável. Em outras palavras, cada linha azul representa um modelo simples que pode ser utilizado para prever $vendas$ utilizando TV, $rádio$ e $jornal$, respectivamente.

Suponha que observamos uma resposta quantitativa Y e p preditores X_1, X_2 e X_p. Assumimos que há algum relacionamento entre Y e $X = (X_1, X_2, ..., X_p)$, o que pode ser escrito na forma geral

$$Y = f(X) + \varepsilon$$

(2.1)

Aqui f é uma função fixa mas desconhecida de $X_1, X_2, ..., X_p$, e ε é um termo de erro aleatório, que é independente de X e possui média zero. Nessa formulação, f representa a informação sistemática que X fornece sobre Y.

Em outro exemplo, considere um conjunto de dados $salário$ e um gráfico de $salário\, x\, anos\, de\, estudo$ para 30 pessoas. O gráfico sugere que é possível prever o salário por meio do número de anos de estudo de uma pessoa. No entanto, a função f que conecta a variável de entrada à variável de saída é em geral desconhecida. Dessa forma, é possivel estimar f com base nos pontos observados. Como "salário" é um conjunto de dados simulado, f é conhecida e representada pela curva azul no painel direito da Figura 2.2. As linhas verticais representam os termos de erro ε. Notamos que algumas das 30 observações estão acima da curva azul e algumas estão abaixo; logo, os erros possue média aproximadamente igual a zero. Em geral, a função f pode envolver mais que uma variável de entrada. Na figura 2.3. colocamos em um gráfico $salário$ como função de $anos\, de\, estudo$ e $prestígio$. Daí, f é uma superfície bidimensional que deve ser estimada com base nos dados observados.

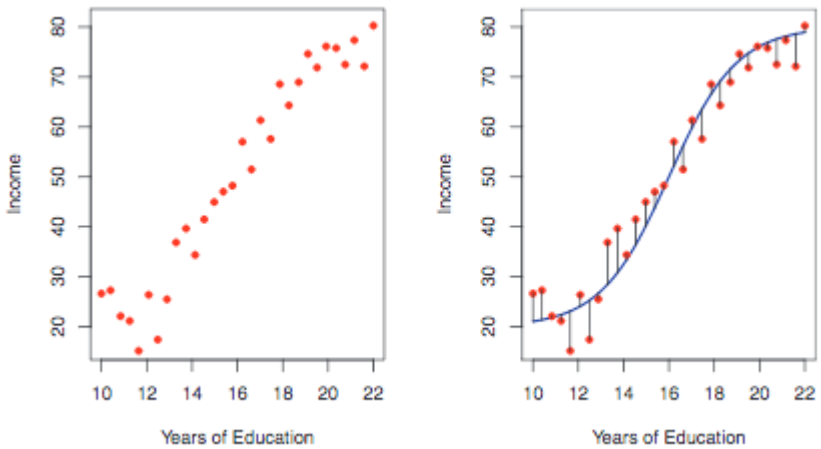

Figura 2.2.
O conjunto de dados $Salário$. À esquerda: os pontos vermelhos são os valores observados de $salário$ (em dezenas de milhares de dólares) e $anos$ de $estudo$ para 30 indivíduos. À direita: A curva azul representa a verdadeira relação oculta entre $salário$ e $anos\, de\, estudo$, a qual é geralmente desconhecida (mas é conhecida neste caso, pois os dados são simulados). As linhas pretas representam o erro associado a cada observação. Note que alguns erros são positivos (acima da curva azul) e alguns são negativos (abaixo da curva). No geral, estes erros possuem média aproximada de zero.

Essencialmente, aprendizado estatístico se refere a um conjunto de abordagens para estimar f.

2.1.1. Por que estimar f?

Há duas razões principais que precisaremos para estimar f: predição e interfêrencia.

Predição

Em muitas situações, um conjunto de entradas X estão disponíveis, mas a saída Y não pode ser facilmente obtida. Neste contexto, como o termo de erro tende a ter média zero, podemos prever Y usando

$$\bar{Y} = \bar{f}(X)$$

(2.2)

onde $\bar{f}$ representa nossa estimativa para f, e $\bar{Y}$ representa a predição resultante para a Y. Desta forma, $\bar{f}$ é frequentemente tratada como uma caixa preta, no sentido em que não estamos tipicamente preocupados com a forma exata de $\bar{f}$, visto que fornece predições corretas para Y.

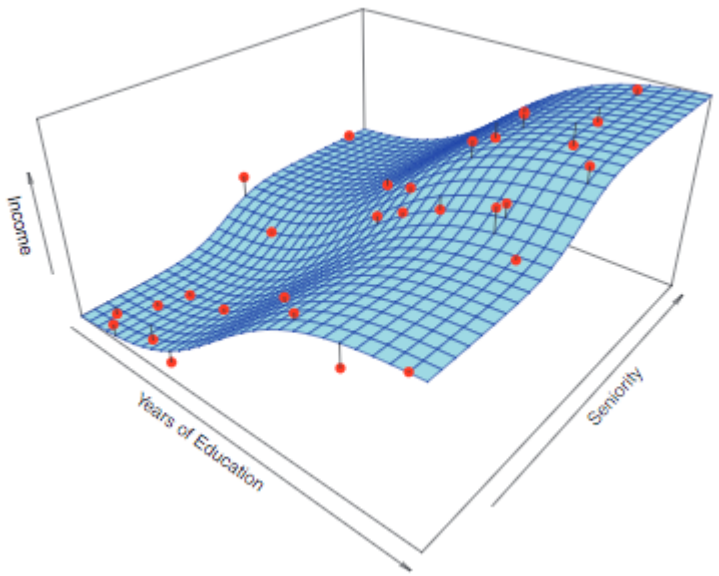

Figura 2.3.
O gráfico apresenta $salário$ *como função de* $anos\ de\ estudo$ *e* $prestígio$ *no conjunto de dados* $Salário$. *A curva azul representa a verdadeira relação oculta entre* $salário$ *e* $anosdeeducação$ *e* $prestígio$, *a qual é conhecida, já que os dados são simulados. Os pontos vermelhos indicam os valores observados destas quantidades para 30 indivíduos.*

Por exemplo, suponha que $X_1,...,X_p$ sejam características de uma amostra de sangue de um paciente que pode ser medida em um laboratório, e Y a varíavel que armazena o risco do paciente reagir negativamente a uma droga em particular. Naturalmente, será procurado prever Y utilizando X, já que podemos evitar a administração da droga em questão para pacientes que possuem alto risco de reagir negativamente (isto é para pacientes com estimativas altas de Y).

A acurácia de $\bar{Y}$ como predição para y depende de duas quantidades, que chamaremos de erro redutível e erro irredutível. Em geral $\bar{f}$ não será uma estimativa perfeita para f, e essa inacurácia introduzirá algum erro. Este erro é redutível porque podemos potencialmente aumentar a acurácia de $\bar{Y}$ usando a técnica de aprendizado estatístico mais adequada para estimar f. No entanto, mesmo que fosse possível formar uma estimativa perfeita para f, tal q ue nossa resposta estimada tivesse a forma $\bar{Y} = f(X)$, nossa predição ainda possuiria erros! Isto acontece porque Y é também uma função de ε, que, por definição, não pode ser prevista utilizando X. Portanto, variabilidade associada a ε também afeta a acurácia de nossas predições. Isto é conhecido como erro irredutível, pois não importa o quão bem estimemos f, não poderemos diminuir o erro introduzido

Considere uma dada estimativa $\bar{f}$ e um conjunto de preditores X, os quais fornecem a predição $\bar{Y} = \bar{f}(X)$. Assumiremos por um momento que ambos $\bar{f}$ e X estão fixados. Daí, é fácil mostrar que

$$E(Y - \bar{Y})^2 = E[f(X) + E - \bar{f}(X)]^2 = [f(X) - \bar{f}(X)]^2 + Var(\varepsilon)$$

(2.3)

$[f(X) - \bar{f}(X)]^2$ **= termo redutível**

$Var(\varepsilon)$ **- termo irredutível**

onde $E(Y - \bar{Y})^2$ representa a média, ou valor esperado, da diferença ao quadrado entre os valores previsto e verdadeiro valor de Y, e $Var(\varepsilon)$ representa a variância associada com o termo de erro ε.

Nosso foco será em técnicas para estimar f com o objetivo de minimizar o erro redutível.

Inferência

Frequentemente, estamos interessados em compreender a forma como Y é afetado enquanto $X_1,...,X_p$ muda. Nesta situação desejamos estimar f, mas nosso objetivo não é necessariamente fazer predições para Y. Nós queremos entender o relacionamento entre X e Y, ou mais especificamente, entender como Y muda como função de $X_1,...,X_p$. Agora $\bar{f}$ não pode ser tratada como uma caixa preta, pois precisamos saber sua forma exata. Neste contexto, podemos estar interessados em responder as seguintes perguntas:

Quais preditores estão associados à resposta?

Frequentemente apenas uma pequena fração dos preditores disponíveis estão substancialmente associados com Y. Identificar os preditores importantes dentre um grande conjunto de possíveis varíaveis pode ser extremamente útil, dependendo da aplicação.

Qual o relacionamento entre a resposta e cada preditor?

Alguns preditores podem possuir uma relação positiva com Y, no senso de que, o aumento do preditor é associado com o aumento dos valores de Y. outros preditores podem possuir uma relação oposta. Dependendo da complexidade de f,o relacionamento entre a resposta e um dado preditor pode também depender dos valores dos outros preditores.

O relacionamento entre Y e cada preditor pode ser resumido adequadamente utilizando uma equação linear, ou o relacionamento é mais complicado?

Historicamente, a maioria dos métodos para estimar f têm tomado uma forma linear. Em algumas situações, tal suposição é coerente ou até desejável. Mas, frequentemente, o verdadeiro relacionamento é mais complicado, no qual um modelo linear pode não fornecer uma representação precisa do relacionamento entre as variáveis de entrada e saída.

Por exemplo, considere uma empresa que está interessada em conduzir uma campanha de marketing. O objetivo é identificar indivíduos que responderão positivamente a uma correspondência, baseando-se em observações de variáveis demográficas medidas para cada indivíduo. Neste caso, as variáveis demográficas servem como preditores, e as respostas à campanha de marketing (positivas ou negativas) servem como resultado. A empresa não está interessada em obter uma compreensão profunda dos relacionamentos entre cada preditor individual e a resposta; em vez disso, a empresa simplesmente deseja um modelo preciso para predizer a resposta utilizando os preditores. Este é um exemplo de modelar com o objetivo de prever.

Pode-se estar interessado em responder perguntas tais como:

Qual mídia contribui para as vendas?

Qual mídia gera o maior aumento nas vendas?

Qual parte do aumento das vendas é associada a um dado aumento nas propagandas de TV?

Esta situação cai no paradigma de inferência. Outro exemplo envolve a modelagem da marca de um produto que o consumidor pode adqurir baseado em variáveis como preço, localização do estoque, descontos, preço competitivo, entre outros. Nesta situação pode-se estar interessado em como cada uma das variáveis afeta a probabilidade de compra. Por exemplo, qual será o efeito nas vendas ao mudar o preço de um produto? Este é um exemplo de modelagem para inferência.

Pode-se modelar um problema com foco em inferência e predição. Por exemplo, no cenário da venda de imóveis, pode-se relacionar os valores de casas a entradas como nível de criminalidade, zoneamento, distância até um rio, qualidade do ar, escolas, níveis de salário dos moradores, entre outros. Neste caso podemos estar interessados em como as variáveis de entrada individuais afetam os preços - ou seja, quanto a mais valerá uma casa se ela tiver vista para um rio? Este é um problema de inferência. Por outro lado, alguém pode estar simplesmente interessado em prever o valor de uma casa dadas suas características: a casa está acima do valor do mercado ou abaixo? Este é um problema de predição.

Dependendo se nosso objetivo final é predição, inferência ou uma combinação de ambos, diferentes métodos para estimar f podem ser apropriados. Por exemplo, modelos lineares são relativamente simples e facilmente interpretados, mas podem conduzir a predições não tão precisas como outras abordagens. No entanto, alguns das abordagens altamente não-lineares que serão discutidas futuramente podem produzir predições de Y com grande acurácia, mas sendo menos interpretáveis, o que torna a inferência mais desafiadora.

2.1.2. Como estimar f?

Sempre assumimos que observamos um conjunto de n pontos diferentes. Estas observações são denominadas "dados de treino" porque utilizaremos estas observações para treinar ou ensinar nosso método a estimar f. Seja x_{ij} a representação do valor do j-ésimo preditor, ou entrada, para a observação i, onde i = 1,2,...,n e j = 1,2,...,p, e seja yi a representação da variável de resposta para a i-ésima observação. Daí, nossos dados de treino consistem em $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$ onde $x_i = (x_{i1},x_{i2},...,x_{ip})^T$.

Nosso objetivo é aplicar um método de aprendizado estatístico aos dados de treino para estimar a função desconhecida f. Em outras palavras, queremos encontrar uma função $\bar{f}$ tal que Y ≅ $\bar{f}$(x) para cada observação (X,Y). Generalizando, a maioria dos métodos de aprendizado estatístico para esta tarefa podem ser caracterizados como paramétricos ou não paramétricos.

Métodos paramétricos

Métodos paramétricos envolvem uma modelagem com dois passos.

1. Primeiramente, fazemos uma suposição sobre a forma de f. Por exemplo, uma suposiçãos simples é que f é linear em X:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p.$$

(2.4)

Este modelo é linear. Uma vez que assumimos que f é linear, o problema de estimar f é simplificado em grane parte. Em vez de estimar uma função inteiramente arbitrária p-dimensional $f(X)$, só é necessário estimar os p+1 coeficientes $\beta_0, \beta_1, ..., \beta_p$.

Depois que um modelo foi selecionado, precisamos de um procedimento que utiliza os dados de treino para treinar ou ajustar o modelo. No caso do modelo linear anterior, precisamos estimar os parâmetros β0,β1,...,βp. Isto é, queremos encontrar valores destes parâmetros tais que

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p.$$

A abordagem mais comum para se ajustar o modelo é chamada de mínimos quadrados. No entanto, esta é uma das diversas formas de ajustar o modelo linear.

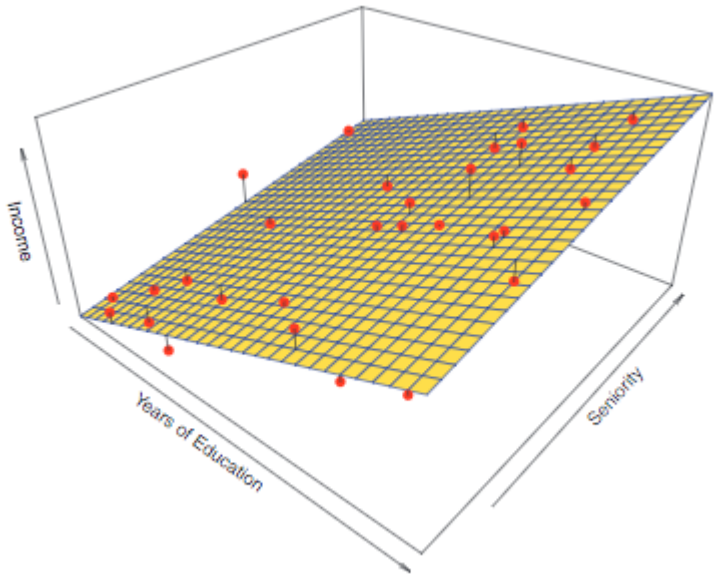

Figura 2.4.
Um ajuste linear por mínimos quadrados para os dados $Salário$ *da figura 2.3. As observações são mostradas em vermelho, e o plano amarelo indica o ajuste por mínimos quadrados aos dados.*

O modelo descrito acima é denominado paramétrico; reduz o problema de estimar f para estimar um conjunto de parâmetros. Assumir uma forma paramétrica para f simplifica o problema de estimar f, pois é geralmente muito mais fácil estimar um conjunto de parâmetros em um modelo linear do que ajustar uma função inteiramente arbritária f. A possível desvantagem de uma abordagem paramétrica é que geralmente o modelo que escolhemos não corresponde à verdadeira forma de f. Se o modelo escolhido for muito longe do verdadeiro f, então nossa estimativa será fraca. Podemos tentar resolver esse problema escolhendo modelos flexíveis que podem se ajustar a diferentes possíveis formas funcionais a f. Mas, em geral, ajustar um modelo flexível requer a estimativa de um número maior de parâmetros. Estes modelos mais complexos podem resultar em um fenômeno conhecido como "overfitting dos dados", o que significa essencialmente que eles seguem os erros muito de perto.

Agora utilizaremos uma abordagem paramétrica aplicada aos dados $salário$. Temos um modelo linear da forma

$$salário = \beta_0 + \beta_1 \, x \, educação + \beta_2 \, x \, prestígio$$

Já que assumimos um relacionamento linear entre a resposta e os dois preditores, o problema se reduz a estimar β_0, β_1 e β_2, nos quais utilizamos a regressão linear por mínimos quadrados. No entanto, a verdadeira função f possui certa curvatura que não é capturada no ajuste linear. Comparando a figura 2.3. à figura 2.4., vemos que o ajuste linear dado na figura 2.4. não é bem verdadeiro: a verdadeira f possui certa curvatura que não é capturada no ajuste lineaar. Mas, o ajuste linear ainda parece razoável ao capturar a relação positiva entre anos de educação e salário, assim como a levemente menor relação entre prestígio e salário. É o melhor que podemos fazer com um pequeno número de observações.

Métodos não-paramétricos

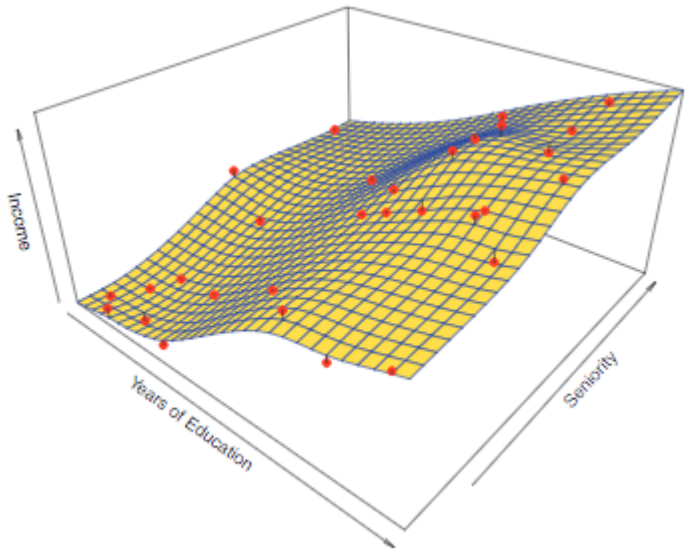

Figura 2.5.
Um ajuste thin-plate spline suave aos dados $Salário$ da figura 2.3. é mostrado em amarelo; as observações estão apresentadas em vermelho.

Métodos não-paramétricos não fazem suposições explícitas sobre a forma funcional de f. Em vez disso, estes procuram uma estimativa de f que chega tão perto quanto o possível dos pontos sem estimativas tão grosseiras. Tais abordagens possuem uma grande vantagem sobre abordagens paramétricas: ao evitar a suposição de uma forma particular para f, elas possuem o potencial de ajustar-se adequadamente a um espectro maior de possíveis formas para f. Qualquer método paramétrico traz a possibilidade da forma funcional utilizada para estimar f ser muito diferente do verdadeiro f, e, consequentemente, o modelo resultante não se ajustará bem aos dados.

Em constraste, abordagens não-paramétricas possuem uma grande desvantagem: como elas não reduzem o problema de estimar f com um pequeno número de parâmetros, são necessárias um grande número de observações (se comparado com o número necessário no método paramétrico) para que se obtenha uma estimativa adequada para f.

Podemos utilizar um método não-paramétrico com os dados $Salário$. Esta abordagem não impõe nenhum modelo pré-especificado em f. Em vez disso, tenta prduzir uma estimativa para f que se aproxime o máximo possível dos dados observados. Neste caso, o ajuste não-paramétrico produziu uma estimativa precisa do verdadeiro f. Pode-se selecionar um grau de suavidade no ajuste, mas deve-se ter cuidado com o "overfitting", pois é uma situação indesejada, já que o ajuste obtido não fornecerá estimativas precisas das respostas para novas observações que não eram parte dos dados de treinos originais.

Um exemplo de abordagem não-paramétrica ao ajustar os dados $Salário$ é mostrado na figura 2.5. Um *thin-plate spline* é utilizado para estimar f. Esta abordagem não impõe nenhum modelo pré-especificado em f. Em vez disso, tenta produzir uma estimativa para f a mais próxima possível dos dados observados, sujeito ao ajuste suave. Neste caso, o método não-paramétrico produziu uma estimativa notavelmente precisa da verdadeira f mostrada na figura 2.3. Para se ajustar um thin plate spline, o analista de dados deve selecionar um nível de suavidade. A figura 2.6. mostra o mesmo ajuste thin-plate spline utilizando um nível menor de suavidade, permitindo um ajuste mais "grosseiro". O resultado da estimativa ajusta os dados observados perfeitamente! No entanto, o ajuste mostrado na figura 2.6. é bem mais variável que a verdadeira função f da figura 2.3. Este é um exemplo de "overfitting" dos dados. É uma situação indesejável porque o ajuste obtido não entregará estimativas precisas da resposta para novas observações (observações que não eram parte original do conjunto de dados de treino).

Portanto, há vantagens e desvantagens ao se utilizar métodos paramétricos e não-paramétricos para aprendizado estatístico.

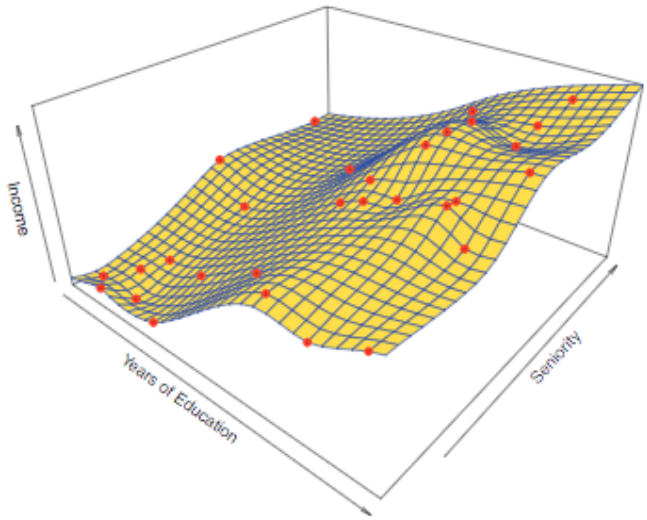

Figura 2.6.
Um ajuste thin-plate spline mais "grosseiro" aos dados $Salário$ da figura 2.3. Este ajuste possui erros nulos nos dados de treino.

2.1.3. A relação de compromisso entre a acurácia da predição e a interpretabilidade de um modelo

Alguns métodos são menos flexíveis ou mais restritos, no sentido de que podem produzir apenas um pequeno espectro de formas para estimar f. Por exemplo, regressão linear é uma abordagem relativamente inflexível, pois pode gerar apenas funções lineares como as linhas mostradas na Figura 2.1. ou o plano mostrado na figura 2.4.

Outros métodos, como *thin-plate splines*, mostrados nas figuras 2.5. e 2.6., são considerados mais flexíveis porque podem gerar um espectro muito maior de possíveis formas de estimar f.

Pode-se perguntar: por que alguém escolheria um método mais restritivo em vez de um mais flexível? Há um grande número de razões para preferirmos um modelo mais restritivo. Se estamos principalmente interessados em inferência, então modelos restritivos são muito mais interpretáveis. Por exemplo, quando o objetivo é inferência, o modelo linear é uma boa opção, pois é fácil de entender a relação entre y e $X_1,X_2,...,X_p$. Modelos muito flexíveis, como os splines mostrados nas figuras 2.5. e 2.6.., e métodos de *boosting* podem levar a estimativas complicadas de f, dificultando o entendimento de como um preditor individual é associado à resposta.

A figura 2.7. fornece uma ilustração da relação de compromisso entre flexibilidade e interpretabilidade para alguns do métodos citados neste livro. Por exemplo, a regressão linear por mínimos quadrados é relativamente inflexível mas é bem interpretável. O lasso depende do modelo linear (2.4), mas utiliza um procedimento de ajuste alternativo para estimar os coeficientes. Este novo procedimento é mais restritivo ao estimar os coeficientes, e alguns são igualados a zero. Desta forma, o lasso é uma abordagem mais flexível que a regressão linear, porque no modelo final a varíavel de resposta será relacionada apenas a um pequeno subconjunto de preditores - aqueles com coeficientes não-nulos.

Modelos aditivos generalizados (GAMs) extendem o modelo linear para permitir certas relações não-lineares. Consequentemente, são mais flexíveis que regressão linear. Eles também são de certa forma menos interpretáveis que regressão linear, pois o relacionamento entre cada preditor e a resposta agora é modelado utilizando uma curva. Finalmente, métodos completamente não-lineares como *bagging*, *boosting* ou *máquinas de vetores de suporte* com núcleos não-lineares são métodos altamente flexíveis e mais difíceis de interpretar.

Estabelecemos que quando o objetivo é a inferência, há claras vantagens em utilizar métodos de aprendizado estatístico simples e relativamente inflexíveis. Em alguns casos, no entanto, estamos interessados apenas em predição, e a interpretabilidade do modelo preditivo não é de interesse. Por exemplo, se procuramos desenvolver um algoritmo para prever o preço de um estoque, nossa única condição é que o algoritmo seja preciso (interpretabilidade não é uma preocupação). Neste contexto, podemos esperar que a melhor abordagem é o modelo mais flexível disponível. No entanto, não é sempre o caso! Frequentemente encontraremos predições mais precisas utilizando um método menos flexível. Este fenômeno, que pode parecer não-intuitivo, é relacionado com o potencial para "overfitting" em métodos altamente flexíveis. Vimos um exemplo disto na figura 2.6.

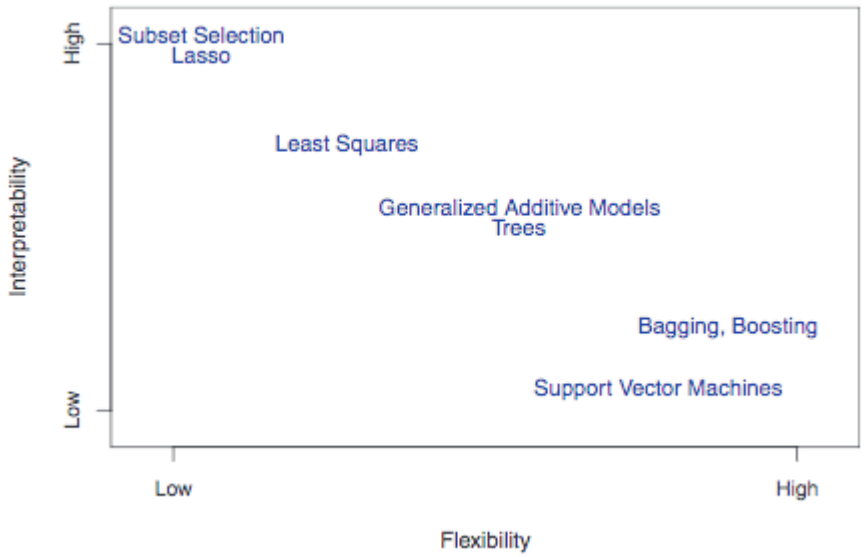

Figura 2.7
A representação da relação de compromisso entre flexibilidade e interpretabilidade, utilizando diferentes métodos de aprendizado estatístico. Em geral, à medida que a flexibilidade de um método aumenta, sua interpretabilidade dimimnui.

2.1.4. Aprendizado supervisionado vs. aprendizado não-supervisionado

A maioria dos problemas de aprendizado estatístico se encaixam em uma das duas categorias: supervisionado ou não-supervisionado. Até agora só discutimos problemas de aprendizado supervisionado. Para cada observação das medidas de predição x_i, i = 1,...,n há uma medida de resposta associada y_i. Desejamos ajustar um modelo que relaciona a resposta aos preditores, com o objetivo de prever a resposta em futuras observações ou obter melhor entendimento da relação entre a resposta e os preditores (inferência). Muitos métodos de aprendizado estatístico clássicos como regressão linear e regressão logística operam no domínio do aprendizado supervisionado.

No entanto, o aprendizado não-supervisionado descreve uma situação mais desafiante, na qual para cada observação i = 1,...,n observamos um vetor de medidas x_i, mas nenhuma resposta associada y_i. Não é possível ajustar um modelo de regressão linear, pois não há variável de resposta para predizer. A situação é denominada não-supervisionada porque não há uma resposta que supervisione nossa análise.

Podemos procurar entender as relações entre as variáveis ou as observações. Uma ferramenta de aprendizado estatístico que podemos utilizar neste contexto é o clustering. O foco do clustering é verificar, com base em $x_1,...,x_n$, se as observações caem em grupos relativamente distintos. Por exemplo, em um estudo de segmentação de mercado podemos observar múltiplas características (variáveis) para potenciais clientes, como CEP, renda familiar e hábitos de compra. Podemos acreditar que os consumidores caem em diferentes grupos, como de grandes gastos versus pequenos gastos. Se as informações sobre os padrões de hábitos de compra estivessem disponíveis, uma análise supervisionada seria possível. No entanto, estas informações não estão disponíveis (não podemos determinar se um cliente gasta muito ou pouco). Neste contexto, podemos tentar agrupar os clientes com base nas variáveis medidas com o objetivo de identificar grupos distintos de clientes em potencial. Identificar tais grupos pode ser de interesse, pois estes podem diferir entre si, como, por exemplo, hábitos de compras.

Ilustraremos um problema de clustering. Podemos colocar em um gráfico 150 observações com medidas em duas varíaveis (X_1 e X_2). Cada observação corresponde a um de três grupos distintos. Pode haver uma sobreposição entre os grupos, e isso torna a tarefa algo mais desafiador. Não é esperado deste método colocar todos os pontos sobrepostos nos seus grupos corretos.

No entanto, pode-se encontrar dados que possuam mais de duas variáveis. Neste caso não podemos colocar em um gráfico as observações tão facilmente. Por exemplo, se há p variáveis no nosso conjunto de dados, então $p(p-1)/2$ gráficos de dispersão distintos podem ser feitos, e então a identificação visual torna-se algo inviável. Por essa razão, métodos automatizados de clustering são importantes.

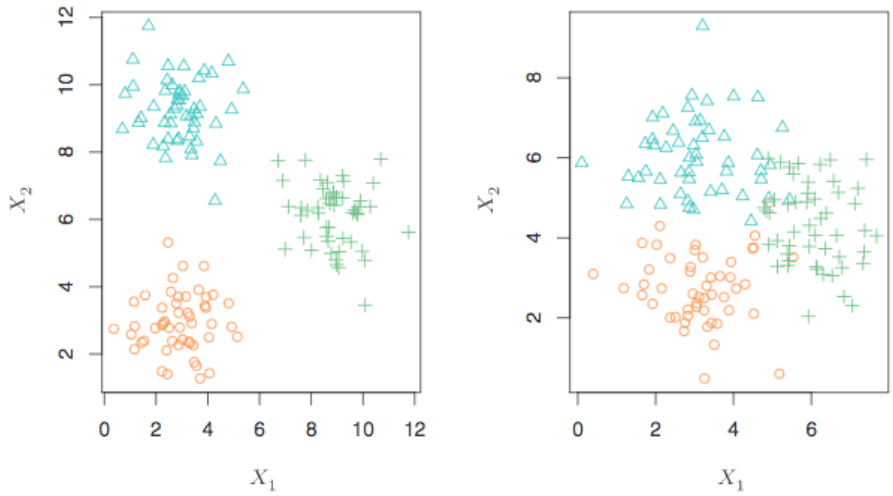

Figura 2.8
Clustering de um conjunto de dados envolvendo três grupos. Cada grupo é representado com um símbolo diferente. À esquerda: Os três grupos estão bem separados. Neste contexto, uma abordagem de clustering identificaria com sucesso os três grupos. À direita: Há sobreposição entre os grupos. Utilizar a abordagem clustering é mais desafiadora nessa situação.

Muitos problemas caem naturalmente nos paradigmas de aprendizado supervisionado ou não-supervisionado. No entanto, às vezes definir se uma análise deve ser considerada supervisionada ou não é mais complicado. Por exemplo, suponha que temos um conjunto de n observações. Para m das observações, onde $m < n$, temos medidas de predição e resposta. Para as $n - m$ observações restantes, temos medidas de predição mas nenhuma medida de resposta. Tal cenário pode aparecer quando as predições podem ser medidas de forma relativamente barata mas as respostas correspondentes são bem mais caras de se obter. Nós nos referimos a este contexto como aprendizado "semi-supervisionado". Neste contexto, desejamos usar um método de aprendizado de máquina que incorpora as m observações para quais as respostas estão disponíveis, mas também incorporam as n-m observações que não possuem resposta.

2.1.5. Problemas de regressão versus problemas de classificação

Variáveis podem ser caracterizadas como qualitativas ou quantitativas. Variáveis quantitativas assumem valores numéricos (idade, altura, salário, etc.). Por outro lado, variáveis qualitativas assumem valores em uma de K classes diferentes, ou categorias (gênero, marca de um produto, diagnóstico de câncer, etc.).

Tendemos a nos referir a problemas com uma resposta quantitativa como problemas de regressão, e a problemas com uma resposta qualitativa como problemas de classificação. Regressão linear por mínimos quadrados é normalmente usada com uma resposta quantitativa, enquanto regressão logística é tipicamente usada com uma resposta qualitativa. Alguns métodos estatísticos, como os k-vizinhos mais próximos e boosting, podem ser utilizados em ambas as respostas.

Nós tendemos a selecionar métodos de aprendizado estatístico com base se a resposta é quantitativa ou qualitativa; por exemplo, podemos utilizar regressão linear quando a resposta é quantitativa e regressão logística quando é qualitativa. No entanto, é considerado menos importante se os preditores são qualitativos ou quantitativos.

2.2. Avaliando a acurácia do modelo

Em estatística, não há um método melhor que todos os outros. Em um conjunto de dados particular, um método específico pode funcionar melhor, mas outro método pode funcionar melhor em um conjunto de dados diferente. Logo, é uma tarefa importante decidir para algum conjunto de dados qual método produz os melhores resultados. A tarefa de selecionar a melhor abordagem pode ser uma das partes mais desafiantes de realizar aprendizado estatístico na prática.

2.2.1. Medindo a qualidade do ajuste

Para avaliar a performance de um método de aprendizado estatístico em um dado conjunto de dados, precisamos de alguma forma para medir o quão bem as predições correspondem aos dados observados. Ou seja, precisamos quantificar até onde a resposta prevista para uma dada observação é próxima ao verdadeiro valor para aquela observação. No contexto de regressão, a medida mais comumente usada é o erro quadrático médio (MSE), dado pela equação

$$MSE = {}^{1}\!/_{n}\Sigma(y_i - \bar{f}(x_i))^2$$

(2.5)

onde $\bar{f}(x_i)$ é a predição que $\bar{f}$ fornece para a i-ésima observação. O erro quadrático médio será pequeno se as respostas previstas estão bastante próximas às verdadeiras respostas, e será grande se para algumas das observações as respostas previstas e verdadeiras diferem substancialmente.

O erro quadrático médio em (2.5.) é computado usando os dados de treino que foram usados para ajustar o modelo, e poderia ser denominado mais corretamente de erro quadrático médio de treino. Mas, em geral, não nos importamos realmente o quão bem o método funciona nos dados de treino. *Estamos interessados na acurácia das predições que obtemos quando aplicamos nosso método para dados de teste não vistos anteriormente.*

Suponha que estamos interessados em desenvolver um algoritmo para prever o preço de um estoque baseado em retornos de estoque anteriores. Mas, não nos importamos o quão bem o método prevê o preço da semana passada, e sim da próxima semana ou do próximo mês. Ou, podemos, ter medidas clínicas (peso, pressão sanguínea, altura, histórico de doença na família) para um número de pacientes, e também temos informações se cada um dos pacientes possui diabetes. Podemos utilizar estes pacientes para treinar um método de aprendizado estatístico para prever o risco de diabetes baseado em medidas clínicas. Na prática, queremos que o método identifique adequadamente o risco de diabetes para futuros pacientes baseados em seus dados. Não estamos muito interessados se o método prevê diabetes nos pacientes usados para treinar o modelo, pois já sabemos qual dos pacientes possui diabetes.

Explicando matematicamente, suponha que ajustamos nosso método de aprendizado estatístico nas nossas observações de treino $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$ e obtemos a estimativa $\bar{f}$. Podemos então computar $\bar{f}(x_1),\bar{f}(x_2),...,\bar{f}(x_n)$. Se estes são aproximadamente iguais a $y_1,y_2,...,y_n$ então o erro quadrático médio de treino será pequeno. No entanto, não estamos muito interessados se $\bar{f}(x_i) \cong y_i$. Em vez disso, queremos saber se $\bar{f}(x_0)$ é aproximadamente igual a y_0, no qual (x_0,y_0) é uma observação de teste não vista antes e nem utilizada para treinar o método de aprendizado estatístico. Devemos escolher o método que retorna o menor erro quadrático médio de treino. Em outras palavras, se tivéssemos um grande número de observações de teste, poderíamos computar

$$Média(y_0 - \bar{f}(x_0))^2$$

(2.6)

a média do erro quadrado de predição para estas observações de teste (x_0,y_0). Gostaríamos de selecionar o modelo para qual a média dessa quantidade - o erro quadrático médio de teste - é o menor possível.

Como selecionamos um método que minimiza o erro quadrático médio de teste? Em alguns contextos, há um conjunto de dados de teste disponível - ou seja, temos acesso a um conjunto de observações que não foram usadas para treinar o método de aprendizado estatístico. Podemos simplesmente avaliar (2.6) nas observações do teste, e selecionar o método de aprendizado para qual o erro quadrático médio de teste é o menor. Mas e se não houver observações de teste? Nesse caso, pode-se imaginar que basta selecionar um método de aprendizado estatístico que minimiza o erro quadrático médio de teste. Parece uma abordagem sensata, já que o erro quadrático médio de teste e o erro quadrático médio de treino são aparentemente parecidos. Infelizmente, há um problema fundamental com esta estratégia: não há garantia de que o método com o menor erro quadrático médio de treino também terá o menor erro quadrático médio de teste. De grosso modo, há muitos métodos estatísticos que estimam especificamente coeficientes que minimizam o conjunto de erros quadráticos médios de treino. Para esses métodos, o MSE de treino pode ser pequeno, mas o MSE de teste é frequentemente muito maior.

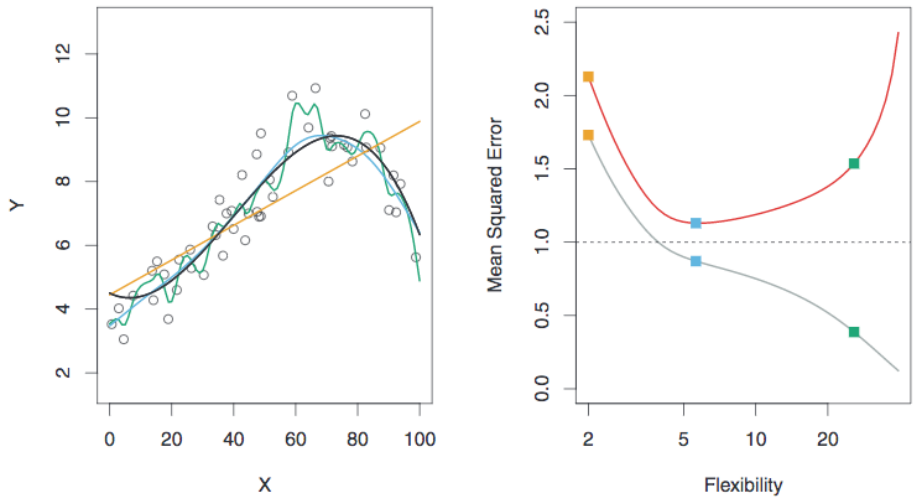

Figura 2.9.
À esquerda: Dados simulados de f, mostrado em preto. Três estimativas de f são mostradas: a linha de regressão linear (curva laranja), e dois ajustes smoothing spline (curvas azul e verde). À direita: MSE de treino (curva cinza), MSE de teste (curva vermelha), e a média mínima possível de erro quadrático sobre todos os métodos (linha quadriculada). Os quadrados representam as MSEs de treino e teste para os três ajuste mostrados no painel à esquerda.

A figura 2.9. ilustra este fenômeno com um exemplo simnples. No painel do lado esquerdo da figura 2.9, foram geradas observações de (2.1) com a verdadeira f dada pela curva preta. As curvas laranja, azul e verde ilustram três estimativas possíveis para f obtidas utilizando métodos com diferentes níveis de flexibilidade. A linha laranja é o ajuste de regressão linear, o qual é relativamente inflexível. As linhas azul e verdes foram produzidas utilizando *smoothing splines*, com diferentes níveis de suavidades. É claro que à medida que o nível de flexibilidade aumenta, o ajuste das curvas se ajustam aos dados observados de forma mais próxima. A curva verde é a mais flexível e se ajusta muito bem aos dados; no entanto, observamos que se ajusta à verdadeira f (mostrada em preto) de forma pobre, pois é muito sinuosa. Ao ajustar o nível de flexibilidade do ajuste da smoothing spline, podemos produzir diferentes ajustes aos dados.

Agora, moveremos para o painel do lado direito da figura 2.9. A curva cinza mostra o MSE médio de treino como função da flexibilidade, ou *graus de liberdade*, para um número de smoothing splines. Grau de liberdade é uma quantidade que resume a flexibilidade de uma curva. Os quadrados laranja, azul e verde indicam os MSEs associados com as curvas correspondentes no painel do lado esquerdo. Uma curva mais suave e mais restrita possui menos graus de liberdade que uma curva sinuosa - note que na figura 2.9, a regressão linear é a mais restrita, com dois graus de liberdade. O MSE de treino declina à medida que a flexibilidade aumenta. Neste exemplo, a verdadeira f é não-linear, e então o ajuste linear laranja não é flexível o suficiente para estimar f bem. A curva verde possui o menor MSE de treino de todos os três métodos, já que corresponde ao ajuste mais flexível entre as três curvas no painel à esquerda.

Neste exemplo, sabemos a verdadeira função f, e então podemos também computar o MSE de teste sobre um grande conjunto de teste, como função da flexibilidade (obviamente, em geral f é desconhecido, então isso não será possível). O MSE de teste é mostrado utilizando a curva vermelha no painel à direita, na figura 2.9. Como com o MSE de treino, o MSE de teste inicialmente declina à medida que o nível de flexibilidade aumenta. No entanto, em algum ponto, os níveis do MSE de teste se nivelam e começam a crescer novamente. Consequentemente, ambas as curvas laranja e verde possuem alto MSE. A curva azul minimiza o MSE de teste, o que não deveria ser surpreendente, já que é a curva que aparentemente melhor estima f no painel à esquerda da figura 2.9. A linha horizontal pontilhada indica Var(ε), o erro irredutível em (2.3), o que corresponde ao menor MSE de teste alcançado entre todos os métodos possíveis. Daí, a smoothing spline representada pela curva azul é próxima à ótima.

No painel à direita da figura 2.9., à medida que a flexibilidade do método de aprendizado estatístico aumenta, observamos uma diminuição monótona no MSE de treino e um *formato em U* no MSE de teste. Esta é uma propriedade fundamental do aprendizado estatístico que permanece independentemente do conjuntos de dados em particular e independentemente do método estatístico sendo utilizado. À medida que a flexibilidade do modelo aumenta, o MSE de

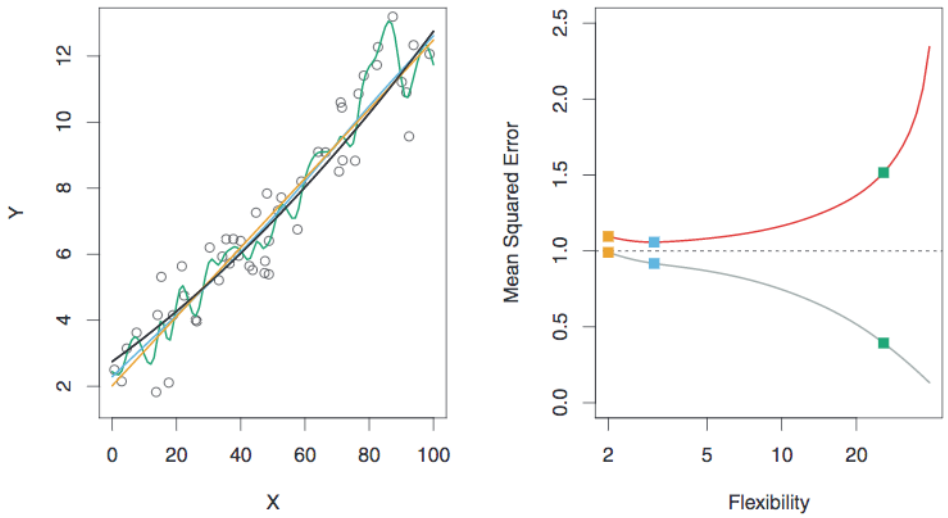

Figura 2.10.
Os detalhes são como os utilizados na figura 2.9., utilizando outra função f verdadeira muito mais próxima a linear. Neste contexto, a regressão linear fornece um ótimo ajuste aos dados.

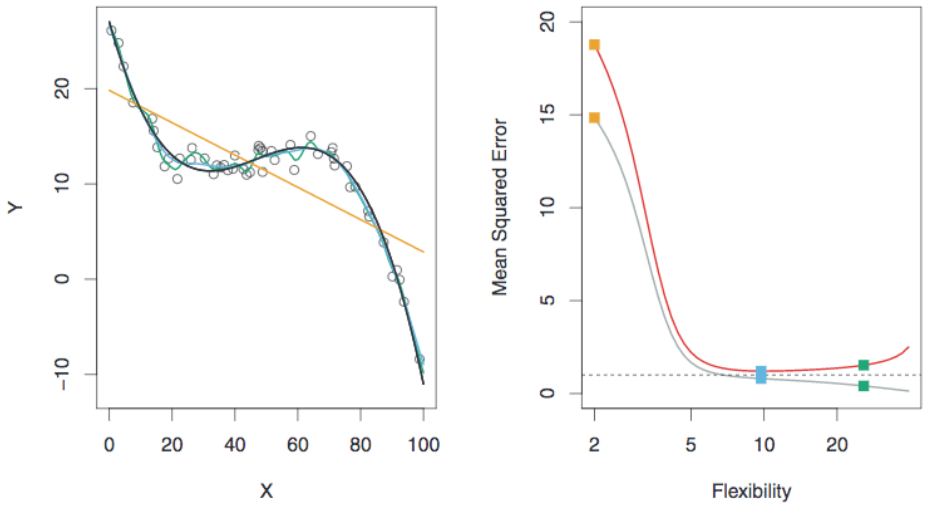

Figura 2.11.
Os detalhes são como os utilizados na figura 2.10., utilizando outra função f bem diferente da linear. Neste contexto, a regressão linear fornece um ajuste pobre aos dados.

A figura 2.10. fornece outro exemplo no qual a verdadeira f é aproximadamente linear. Novamente, observamos que o MSE de treino diminui à medida que a flexibilidade aumenta, e que há um *formato em U* no MSE de teste. No entanto, porque a verdadeira forma de f é próxima à linear, o MSE de teste diminui levemente antes de aumentar novamente, tal que o ajuste por mínimos quadrados laranja é substancialmente melhor que a curva verde altamente flexível. Finalmente, a figura 2.11. mostra um exemplo no qual f é altamente não-linear. As curvas do MSE de treino e do MSE de teste ainda exibem os mesmos padrões gerais, mas agora há uma rápida diminuição em ambas as curvas antes que o MSE de teste comece a aumentar vagarosamente.

Na prática, geralmente se pode computar o MSE de treino com relativa facilidade, mas estimar o MSE de teste é consideravelmente mais difícil, pois geralmente não há dados de teste disponíveis. Como os três exemplos anteriores ilustram, o nível de flexibilidade correspondente ao modelo com o MSE de teste mínimo pode variar consideravelmente ao longo de conjuntos de dados.

2.2.2. O "Trade-off" entre variância e polarização

A forma em U observada nas curvas do MSE de teste (Figuras 2.9. a 2.11.) é o resultado de duas propriedades dos métodos de aprendizado estatístico. É possivel mostrar que o MSE de teste esperado, para um dado valor x_0, pode ser sempre decomposto em uma soma de três valores fundamentais: a variância de $\bar{f}(x_0)$, a polarização ao quadrado de $\bar{f}(x_0)$ e a variância do termo de erro ε. Ou seja,

$$E(y_0 - \bar{f}(x_0))^2 = Var(\bar{f}(x_0)) + [Bias(\bar{f}(x_0))]^2 + Var(\varepsilon).$$

(2.7)

A notação $E(y_0 - \bar{f}(x_0))^2$ define o MSE de teste esperado, e se refere à MSE de teste média que obteríamos se estimássemos f repetidamente usando um grande número de conjuntos de treino, e testássemos cada um em x_0. O MSE de teste esperado, em geral, pode ser computado calculando $E(y_0 - \bar{f}(x_0))^2$ sobre todos os valores possíveis de x_0 no conjunto de teste.

A equação (2.7) nos diz que, para minimizar o erro de teste esperado, precisamos selecionar um método de aprendizado estatístico que simultaneamente alcança *baixa variância* e *baixa polarização*. A variância é inerentemente uma quantidade não-negativa, e a polarização ao quadrado também é não-negativa. Daí, observamos que o MSE de teste esperado nunca pode ficar abaixo de $Var(\varepsilon)$, o erro irredutível de (2.3).

O que queremos dizer por variância e polarização de um método de aprendizado estatístico? Variância se refere à quantidade pela qual $\bar{f}$ mudaria se a estimássemos utilizando um conjunto de dados de treino diferente. Já que os dados de treinos são utilizados para ajustar o método de aprendizado estatístico, conjuntos de dados de treino diferentes resultarão em um $\bar{f}$ diferente. Mas, de preferência, a estimativa para f não deve variar demais entre os conjuntos de treino. No entanto, se um método possui variância alta então pequenas mudanças nos dados de treino podem resultar em grandes mudanças em $\bar{f}$. Em geral, metódos estatísticos mais flexíveis possuem maior variância.

Considere as curvas verde e laranja na figura 2.9. A curva verde flexível está seguindo as observações de perto e possui alta variância porque a mudança de qualquer um destes pontos dos dados pode fazer a estimativa $\bar{f}$ mudar consideravelmente. Em contraste, a linha de mínimos quadrados laranja é relativamente inflexível e possui baixa variância, pois a mudança de qualquer observação provavelmente causará apenas um pequeno deslocamento na posilão da linha.

Por outro lado, polarização se refere ao erro que é introduzido ao aproximar um problema real, o qual pode ser extremamente complicado, por um modelo muito mais simples. Por exemplo, a regressão linear assume que há um relacionamento linear entre Y e $X_1, X_2, ..., X_p$. É improvável que algum problema real possua uma relação tão simplesmente linear, logo, realizar uma regressão linear certamente resultará em alguma polarização na estimativa de f. Na figura 2.11., a verdadeira f é substancialmente não-linear, então, não importa quantas observações de treino são dadas, não será possível produzir uma estimativa precisa utilizando regressão linear. No entanto, na figura 2.10., a verdadeira f é aproximadamente linear, e, com dados suficientes, deve ser possível produzir uma estimativa precisa utilizando regressão linear. Em geral, métodos mais flexíveis resultam em menor polarização.

Como regra geral, à medida que utilizamos métodos mais flexíveis, a variância irá aumentar e a polarização diminuir. A proporção relativa de mudança entre estas duas quantidades determinam se o MSE de teste aumenta ou diminui. À medida

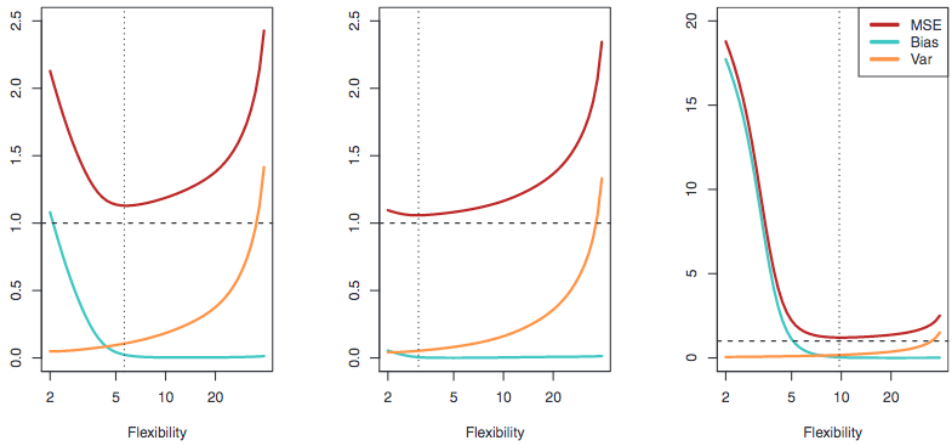

Figura 2.12.
Polarização quadrática (curva azul), variância (curva laranja), Var(ε) (linha quadriculada), e MSE de teste (curva vermelha) para os três conjuntos de dados nas figuras 2.9. a 2.11. A linha vertical pontilhada indica o nível de flexibilidade correspondente ao menor MSE de teste.

2.2.3. O contexto de classificação

Até agora, nossa discussão sobre a acurácia do modelo foi focada no contexto de regressão. Mas, muitos dos conceitos que utilizamos, como o trade-off entre variância e polarização, transferem-se para o contexto da classificação com apenas algumas modificações, já que y_i não é mais numérico. Suponha que queremos estimar f com base nas observações de treino $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$, onde agora $y_1,...,y_n$ são qualitativas. A abordagem mais comum para quantificar a acurácia de nossa estimativa $\bar{f}$ é a $taxa\,de\,erro$ de treino, a proporção de erros que são feitas ao aplicarmos nossa estimativa $\bar{f}$ para as observações de treino:

$$1/n \sum I(y_i \neq \bar{y}_i).$$

(2.8)

Aqui, $\bar{y}_i$ é a classe prevista para a i-ésima observação utilizando $\bar{f}$. E, $I(y_i \neq \bar{y}_i)$ é uma $variável\ indicadora$ que é igual a 1 se $y_i \neq \bar{y}_i$ e zero se $y_i = \bar{y}_i$. Se $I(y_i \neq \bar{y}_i) = 0$ então a i-ésima observação foi classificada corretamente pelo nosso método de classificação; de outra forma, ela foi mal classificada. Portanto, a equação acima computa a fração de classificações incorretas.

A equação 2.8. é denominada taxa de $erro\ de\ treino$ porque é computada com base nos dados que foram utilizados para treinar nosso classificador. Como no contexto de regressão, estamos mais interessados nas taxas de erro que resultam da aplicação do nosso classificador para observações de testes que não foram utilizadas no treino. A taxa de $erro\ de\ teste$ associada A um conjunto de observações de teste da forma (x0,y0) é dada por

$$Média(I(y_0 \neq \bar{y}_0))$$

(2.9)

onde $\bar{y}_0$ é a classe prevista que resulta da aplicação do classificador à observação de teste com preditor x_0. Um bom preditor é o que possui o menor erro de teste (2.9) possível.

O classificador de Bayes

É possível mostrar que a taxa de erro de teste (2.9) é minimizada, em média, por um classificador muito simples que atribui cada observação à classe mais provável, dados seus valores previstos. Em outras palavras, devemos simplesmente atribuir uma observação de teste com um vetor preditor x_0 à classe j para a qual:

$$Pr(Y = j | X = x_0)$$

(2.10)

é maior. Note que a equação acima é uma probabilidade condicional: é a probabilidade de $Y = j$, dado o vetor preditor observado x_0. Este classificador muito simples é denominado "Classificador de Bayes". Em um problema de duas classes em que há apenas dois valores de resposta possíveis (classe 1 ou classe 2), o classificador de Bayes corresponde à classe prevista 1 se (Pr Y =1| X = x_0) > 0.5, e a classe 2 se o contrário.

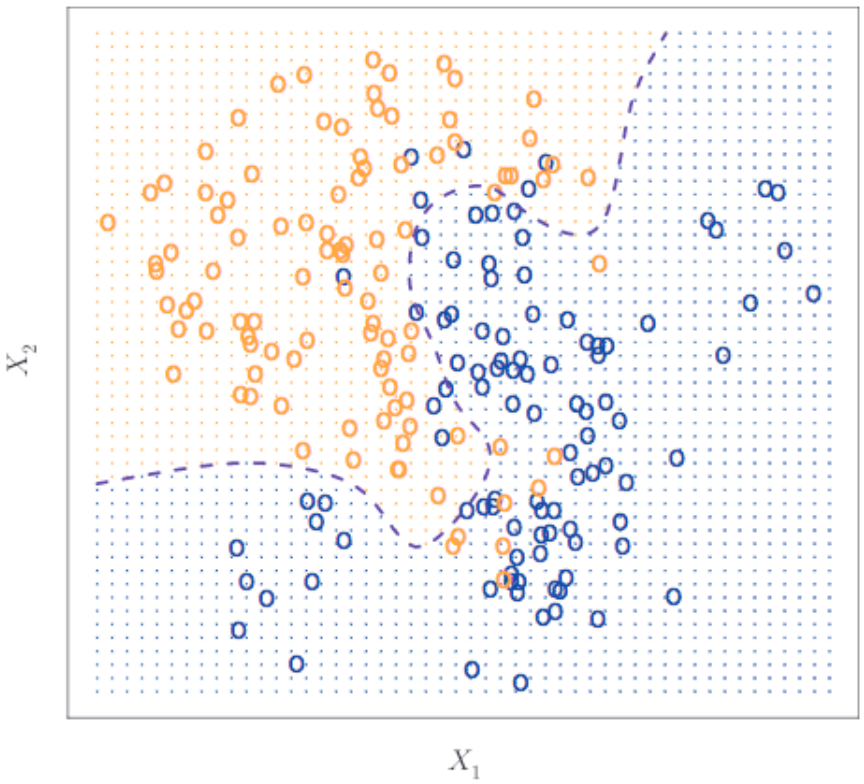

Figura 2.13.
Um conjunto de dados simulado, consistindo de 100 observações em cada um dos dois grupos, indicados em azul e laranja. A linha pontilhada roxa representa a fronteira de decisão de Bayes. A grade de fundo laranja indica a região para qual uma observação de teste será atribuida à classe laranja, e a grade de fundo azul indica a região para qual uma observação de teste será atribuída à classe azul.

A figura 2.13. fornece um exemplo utilizando um conjunto de dados simulado em

K-vizinhos mais próximos

Em teoria, sempre gostaríamos de prever respostas qualitativas utilizando o classificador de Bayes. Mas, para dados reais, não sabemos a distribuição condicional de Y dado X, então computar o classificador de Bayes é impossível. Portanto, o classificador de Bayes serve de padrão inalcançável se comparado aos outros métodos. Muitas abordagens tentam estimar a distribuição condicional de Y dado X, e então classificam uma dada observação à classe com maior probabilidade $estimada$. Um destes métodos é o classificador $K - vizinhos\ mais\ próximos$ (KNN).

Dado um inteiro positivo K e uma observação de teste x_0, o classificador KNN identifica primeiramente os K pontos nos dados de treino que estão mais próximos de x_0, representados por N_0. Então, estima a probabilidade condicional para a classe j como fração dos pontos em N_0 cujos valores de resposta são igual a j:

$$Pr(Y = j | X = x_0) = \frac{1}{K} \Sigma I(y_i = j).$$

(2.12)

Finalmente, o KNN aplica a regra de Bayes e classifica a observação de teste x_0 à classe com a probabilidade mais alta.

A figura 2.14. fornece um exemplo ilustrativo da abordagem KNN. No painel à esquerda, colocamos em um gráfico um pequeno conjunto de dados pequeno consistindo em seis observações laranja e seis observações azul. Nosso objetivo é fazer uma predição para o ponto rotulado pela cruz preta. Suponha que escolhamos K = 3. O KNN identificará primeiramente as três observações que estão mais próximas à cruz. Esta vizinhança é mostrada como um círculo, consistindo de dois pontos azuis e um ponto laranja, resultando em probabilidades estimadas de $\frac{2}{3}$ para a classe azul e $\frac{1}{3}$ para a classe laranja. Consequentemente o KNN irá prever que a cruz preta pertence à classe azul. No painel à direita da figura 2.14. aplicamos a abordagem KNN com K = 3 em todos os valores possíveis para X_1 e X_2, e desenhamos na fronteira de decisão KNN correspondente.

Além do fato de que é uma abordagem bastante simples, o KNN pode frequentemente produzir classificadores que estão surpreendentemente próximos ao classificador de Bayes ótimo. A figura 2.15. mostra a fronteira de decisão KNN, usando K = 10, quando aplicada ao conjunto de dados simulado maior da figura 2.13. Note que mesmo que a verdadeira distribuição não seja conhecida pelo classificador KNN, a fronteira de decisão KNN está muito próxmima da do classificador de Bayes. A taxa de erro de teste utilizando KNN é 0,1363, a qual é próxima à taxa de erro de Bayes de 0,1304.

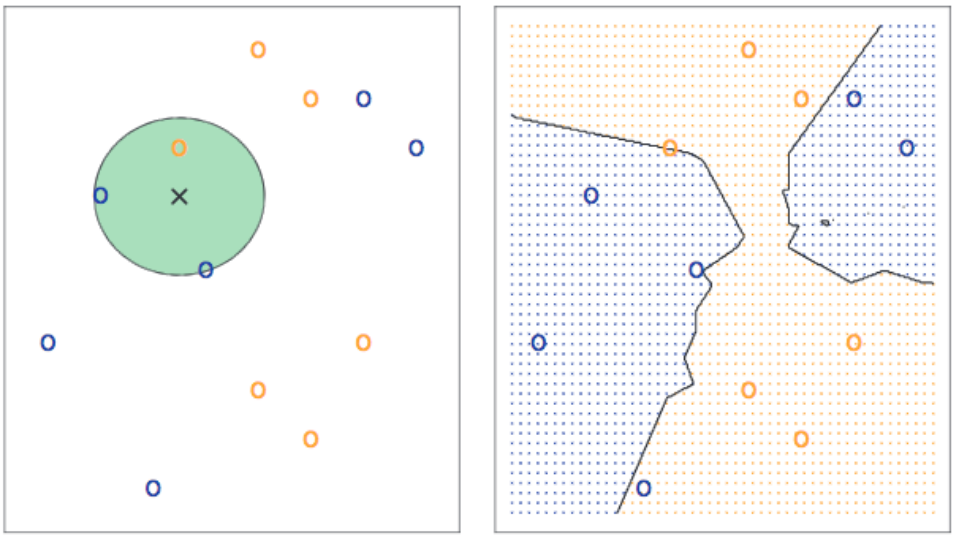

Figura 2.14.
A abordagem KNN, usando K = 3, é ilustrada em uma simples situação com seis observações azuis e seis observações laranja. À esquerda: uma observação de teste na qual um legenda de uma classe prevista é desejada é mostrada como uma cruz preta. Os três pontos mais próximos às observações de teste são identificados, e é previsto que a observação de teste pertence à classe mais recorrente, neste caso a azul. À direita: a fronteira de decisão KNN para este exemplo é mostrada em preto. A grade azul indica a região na qual uma observação de teste será atribuída à classe azul, e a grade laranja indica a região na qual será atribuída à classe laranja.

A escolha de K exerce um efeito drástico no classificador KNN obtido. A figura 2.16. mostra dois ajustes KNN aos dados simulados da figura 2.13., utilizando K = 1 e K = 100. Quando K = 1, a fronteira de decisão é excessivamente flexível e encontra padrões nos dados que não correspondem à fronteira de decisão de Bayes. Isto corresponde a um classificador que possui baixa polarização e variância muito alta. À medida que K aumenta, o método se torna menos flexível e produz uma fronteira de decisão que é próxima à linear. Isto corresponde a um classificador com baixa variância e alta polarização. Neste conjunto de dados simulado, nem K = 1 e K = 100 dão boas predições: eles possuem taxas de erro de teste de 0,1695 e 0,1925, respectivamente.

Como no contexto da regressão, não há uma forte relação entre a taxa de erro de treino e a taxa de erro de teste. Com K = 1, a taxa de erro de treino é 0, mas a taxa de erro de teste pode ser bastante alta. Em geral, à medida que usamos métodos de classificação mais flexíveis, a taxa de erro de treino irá declinar, mas a taxa de erro de teste não. Na figura 2.17., colocamos em um gráfico o KNN de teste e os erros de treino como função de $1/K$. À medida que $1/K$ aumenta, o método se torna mais flexível. Neste contexto de regressão, a taxa de erro de treino declina à medida que a flexibilidade aumenta. No entanto, o erro de teste exibe uma forma de U característica, declinando primeiramente (com um mínimo em aproximadamente K = 10) antes de aumentar novamente, quando o método se torna excessivamente flexível e ocorre um sobreajuste (overfitting).

Em ambos os contextos de classificação e regressão, escolher o nível correto de flexibilidade é crítico para o sucesso de qualquer método de aprendizado estatístico. O trade-off entre polarização e variância, e a forma em U resultante no erro de teste, pode tornar isto uma tarefa difícil.

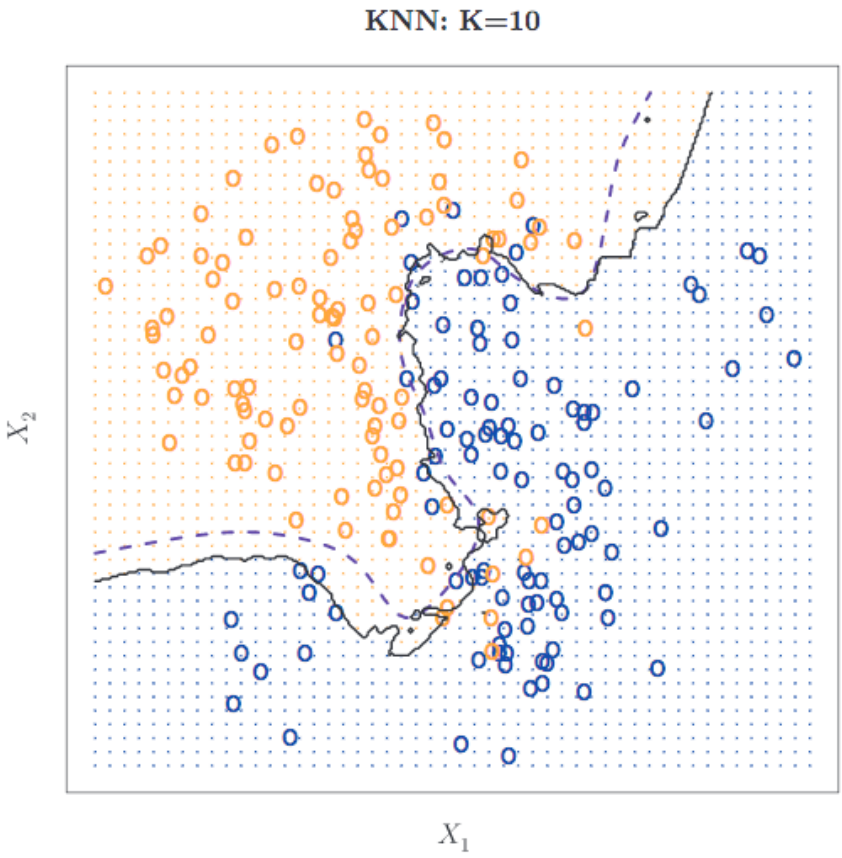

Figura 2.15.
A curva preta indica a fronteira de decisão KNN nos dados da figura 2.13., usando K = 10. A fronteira de decisão de Bayes é mostrada como uma linha pontilhada roxa. As fronteiras de decisão de Bayes e KNN são bastantes similares.

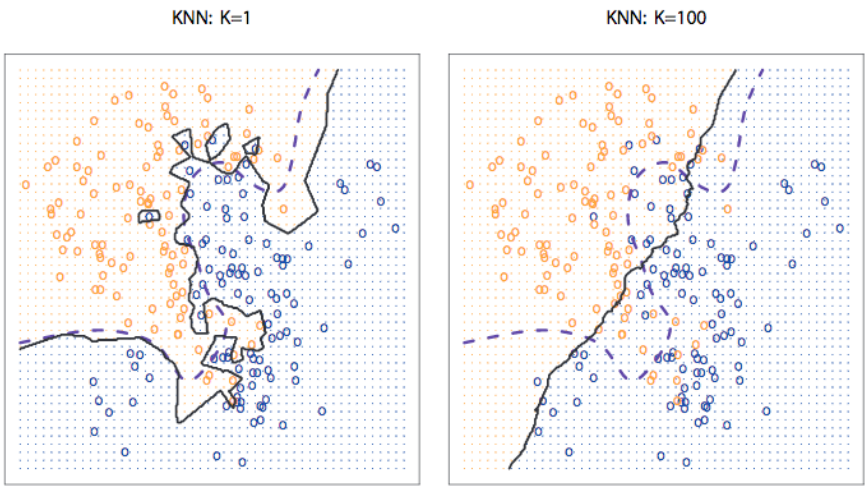

Figura 2.16.
Uma comparação da fronteira de decisão KNN (linhas sólidas pretas) obtidas usando K = 1 e K = 100 nos dados da figura 2.13. Com K = 1, a fronteira de decisão é excessivamente flexível, enquanto com K = 100 não é suficientemente flexível. A fronteira de decisão de Bayes é mostrada como uma linha pontilhada roxa.

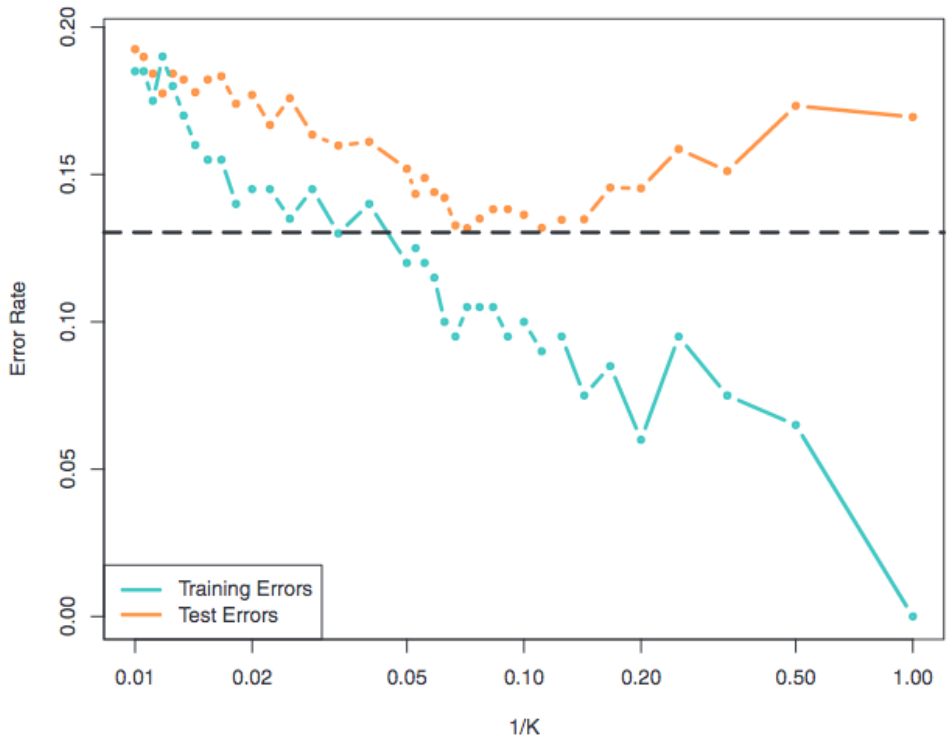

Figura 2.17.
A taxa de erro de treino KNN (azul, 200 observações) e a taxa de erro de teste (laranja, 5000 observações) nos dados da figura 2.13., à medida que o nível de flexibilidade (avaliada usando $1/K$) aumenta, ou equivalentemente à medida que o número de vizinhos K diminui. A linha preta pontilhada indica a taxa de erro de Bayes. Os "saltos" das curvas acontecem devido ao tamanho pequeno do conjunto de dados de treino.