Análise de Regressão

Professor Thiago de Paulo Faleiros

Brasília, 19 de novembro de 2018

- Variância: A variância de n medidas $y_1, y_2, \ldots, y_n$ é definido como

$$\sigma^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n}$$

- Desvio padrão: O desvio padrão de um conjunto de medidas é igual ao quadrado da variância.

Introdução – Análise de Regressão

- Análise de regressão linear significa "ajustar uma reta nos dados"
  - Também chamado modelo linear
- É uma técnica amplamente usada para auxiliar na modelagem e entendimento de problemas do mundo real.
  - fácil de usar
  - fácil de entender intuitivamente
- permite predições

- Análise de Regressão – Método estatístico aplicado na investigação e modelagem da relação entre variáveis
- Mas especificamente, a análise de regressão explora a distribuição de uma variável (ou de algum parâmetro de sua distribuição, como a média), condicionada aos valores de outras variáveis.

- Algumas possíveis aplicações de análise de regressão
  - Existe relação entre a dose administrada de certa medicação e a reduação da pressão arterial?
  - Existe relação entre a nota obtida num exame e o tempo dedicado ao estudo?
  - Qual a relação entre o valor de venda de terrenos em certas localidades com as respectivas áreas?
  -
- A análise de regressão baseia-se na proposição (e ajuste, a partir dos dados amostrados) de funções que expliquem adequadamente a relação entre as variáveis.

- Existe relação da pontuação final de times de um campeonato de futebol com o investimento em contratações e o número de títulos obtidos anteriormente?
- Qual relação do índice de massa corporal de crianças de certas localidades e as seguintes variáveis: renda familiar per-capita, escolaridade da mãe, idade e peso da criança aos nascer?
- Qual a relação entre a quantia aplicada em fundos de determinado banco e características dos clientes como: sexo, ocupação, renda, idade, nível de escolaidade, estado civil, se o cliente tem conta em outros bancos, ...?

Objetivos principais da análise de regressão

- Analisar (descrever) a relação entre uma variável de interesse e uma ou mais variáveis explicativas;
- Retomando o exemplo da pressão arterial vs dose da medicação, para fins de ilustração:
  - A pressão sanguínea diminui conforme se aumenta a dose da medicação? Mantém-se constante? Aumenta?
  - A diminuição na pressão arterial é linear conforme o aumento da dose (diminui a uma taxa constante)? Diminui de forma não linear?
  - Há alguma dose a partir da qual a pressão sanguínea já não responde mais a incrementos na dose? Em algum momento o aumento na dose pode ocasionar o efeito contrário (aumento na pressão)?

Seleção e análise das variáveis que de fato estão relacionadas à resposta

- No estudo do desempenho acadêmico de alunos de certo nível, pode-se ter interesse em identificar variáveis sócio-econômicas e demográficas (como renda familiar, ocupação, escolaridade e situação conjugal dos pais, números de irmãos, tipo de residência, ...) relacionadas;
- No estudo do valor devido por inadimplentes de uma instituição de crédito, pode-se ter interesse em identificar variáveis sócio-econômicas e demográficas (como renda, ocupação, escolaridade, número de filhos, sexo, idade,...), e comportamentais (existência de dívidas anteriores, situação do cliente em diferentes cadastros, como Serasa,...) relacionadas

Estimação de parâmetros e testes de hipóteses de interesse

- A redução na pressão arterial é estatisticamente significativa frente ao aumento na dose da medicação?
- Qual a alteração esperada na pressão arterial decorrente do acréscimo de 1mg na dose da medicação?
- Qual a alteração esperada na pressão arterial decorrente do acréscimo de kmg na dose da medicação?

Predição

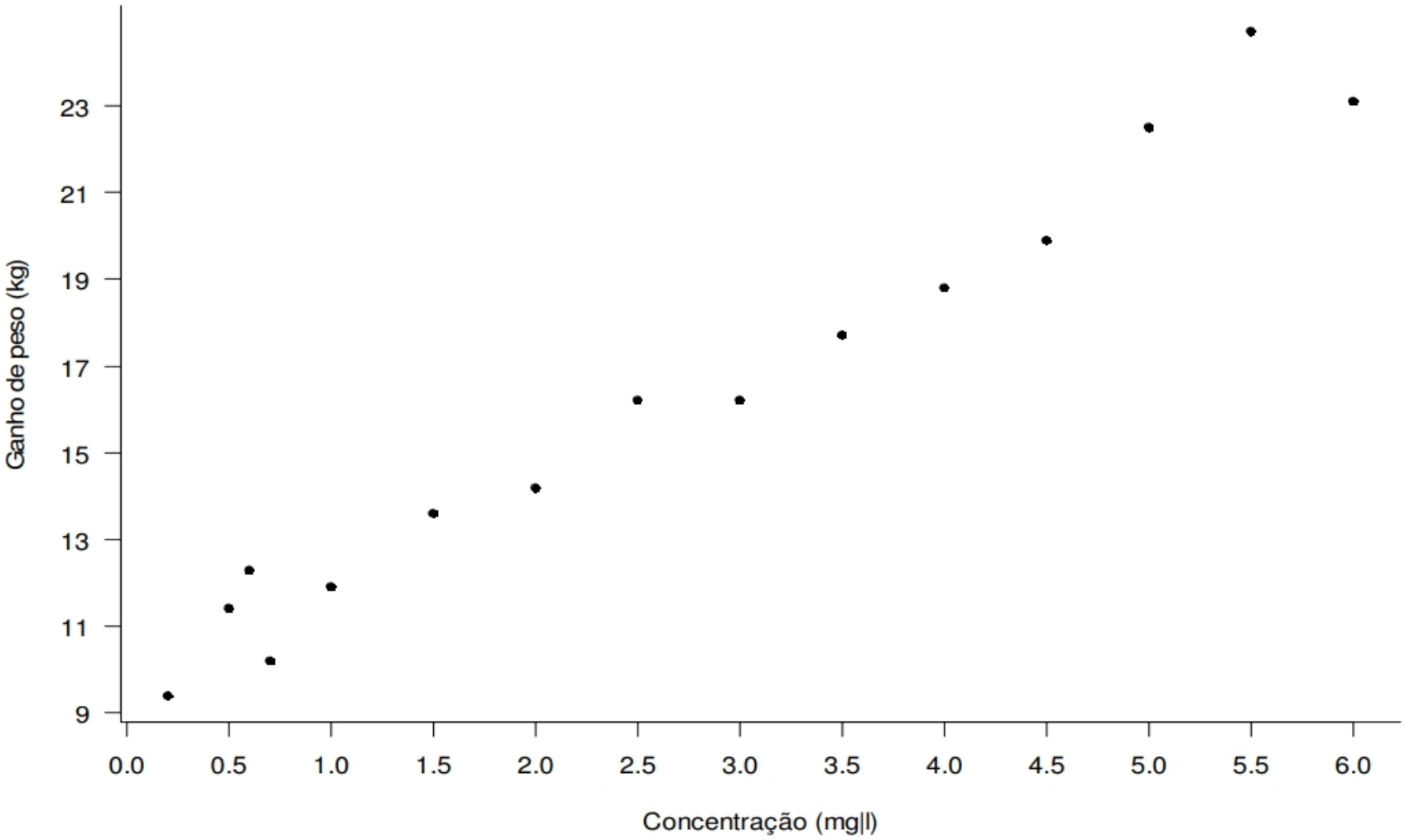
- Qual a diminuição média na pressão arterial para uma dose adminstrada x_o?
- Qual a diminuição a ser observada na pressão arterial para uma dose administrada x_0?
- Qual a dose necessária x_0 para se obter uma resposta desejada y_0?

Exemplo – Deseja-se investigar se o ganho de peso de bovinos está relacionado à quantidade de certa substâncias presente no pasto. Para isso, um estudo foi conduzido com 15 bois de mesma raça e idade, submetidos a dietas com diferentes concentrações da referida substância.
Variáveis consideradas:

- X – Concentração da substância (em mg | litro)
- Y – Ganho de peso do animal após 30 dias (em kg)

Quadro 1 – Concentração da substância adicionada à dieta e ganhos de peso para os 15 bois.

Animal	X	Y	Animal	X	Y
1	0,2	9,4	9	3,0	16,2
2	0,5	11,4	10	3,5	17,7
3	0,6	12,3	11	4,0	18,8
4	0,7	10,2	12	4,5	19,9
5	1,0	11,9	13	5,0	22,5
6	1,5	13,6	14	5,5	24,7
7	2,0	14,2	15	6,0	23,1
8	2,5	16,2			


Ganho de peso (kg)
9
11
13
15
17
19
21
23
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Concentração (mg|l)

- A figura anterior evidencia fortemente uma relação linear entre o ganho de peso e a concentração da substância na dieta
- Uma forma de explificar a relação entre o ganho de peso e a concentração da substância na dieta por meio da equação da reta (modelo) que descreve tal relação, de tal forma que, para um animal submetido a uma dieta com concentração x o ganho de peso fosse dado por:

$$y = \beta_0 + \beta_1 x$$

Note que o ganho de peso não pode ser determinado simplesmente a partir da concentração da substância na dieta (há uma oscilação dos pontos em torno da reta, conforme notado na Figura)

- Assim, um modelo mais apropriado para o problema seria da seguinte forma:

$$y = \beta_0 + \beta_1 x + \epsilon$$

de tal forma que ϵ corresponde à diferença entre o valor observado y e o valor verificado na reta para o respectivo $x(\beta_0 + \beta_1 x)$, $\epsilon = y - (\beta_0 + \beta_1 x)$ configurando uma quantidade aleatória à qual denominamos erro.

- O modelo apresentado em (2) configura um modelo de regressão, e com algumas suposições adicionais acerca da distribuição do componente aleatório que estudaremos adiante, é denominado modelo de regressão linear simples (o termo simples refere-se ao fato dele conter apenas uma variável explicativa).
- No contexto de análise de regressão, é usual denominarmos a variável a ser explicada, como variável resposta (ou variável dependente), e as variáveis que explicam a resposta como variável explicativas (ou variáveis independentes).
- As constantes que compõem o modelo (no caso do exemplo 2 os coeficientes da reta, β_0 e β_1), são denominados parâmetros. Os parâmetros exprimem a relação entre as variáveis.
- O termo ajuste de um modelo refere-se ao processo de estimação dos parâmetros (e, consequentemente, da função de regressão) do modelo com base nos dados disponíveis, ou à própria função de regressão gerada pelas estimativas obtidas.

Poblema de Regressão

- Um problema de regressão é composto de
  - uma variável de resposta Y
  - um número de fatores de riscos da variável predita X_i que afeta Y
    - também chamada variável explanatória, ou características (*features*).
  - uma questão sobre Y – *Como predizer Y sobre diferentes condições?*.
- Y é algumas vezes chamada variável dependente e X_i é variável independente
  - não no mesmo significado de independência estatística
  - configuração experimental onde X_i pode ser modificada em Y pode ser observado

Análise de Regressão: Objetivos

Predição

Nós queremos predizer Y dado valores específicos de X_i.

Inferência de Modelo

Queremos aprender sobre as relações entre as variáveis Y e X_i, tal como as relações das variáveis preditoras que tem mais efeito sobre Y.

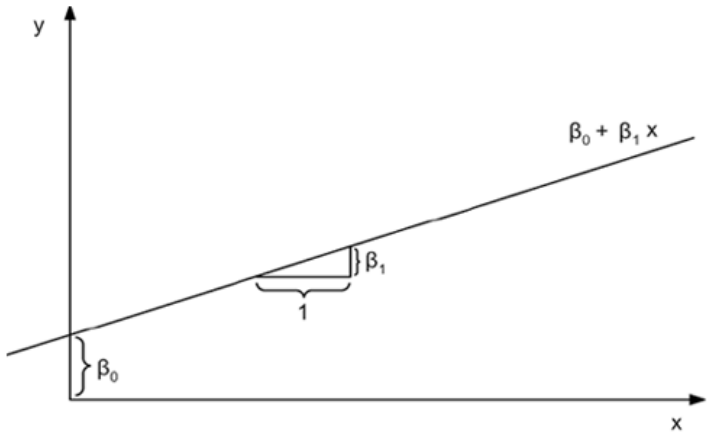
Regressão linear simples

Regressão Linear

- Regressão Linear: um das simples e mais comum técnica de modelagem estatística
- Faz fortes suposições sobre a relação entre as features (variáveis preditoras) X_i e de respostas Y_i.
  - Uma relação linear, plotando uma linha reta entre os pontos
  - Apenas válida para váriáveis de respostas contínuas (não aplicada para respostas categóricas)

$$Y_i = \beta_0 + \beta_1 * X_i + erro, \qquad \text{para i=1 até n} \qquad (1)$$

onde, Y é a variável de resposta, β_0 é o coeficiente linear (intecepto), β_1 é o coeficiente angular (ou coeficiente de regressão, ou a inclinação da reta), e a variável X representa o valor da variável explicativa (variável independente, variável regressora) na i-ésima observação;



- Suponha $Y = \beta_0 + \beta_1 * X + erro$
- Nossa tarefa é estimar β_0 e β_1 considerando os dados disponíveis
- O modelo resultante é a estimativa $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * X + erro$
  - $\hat{Y}$ é a estimativa de Y
- Os valores de β_0 e β_1 são os parâmetros ou coeficientes
- **Objetivo:** minimizar o erro, a diferença entre nossa observação e a predição feita pelo modelo linear induzido

Regressão por mínimos quadrados

- Método para selecionar os parâmetros β_0 e β_1 do modelo
  - Os valores de β_1 e β_0 são escolhidos para minimizar a distância quadrada entre os valores preditos e os valores atuais.
- Aplicação de uma função de custo quadrática.
  - Mapeia uma observação para uma função de custo

- Qual escolha do modelo?
- O que acontece quando $\beta_1 = 0$?
- Queremos saber a contribuição da variável regressora X para explicar a resposta Y, uma vez que se H_0 for verdadeira, isto é, se $B_1 = 0$, essa contribuição não é significativa.

- Variância:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

- Soma de quadrado total – $SQT = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Soma de quadrado do erro – $SQE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Soma de quadrados de regressão – $SQR_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

$$SQT = SQE + SQR_{reg}$$

Tabela Anova

fonte de variação	gl	SQ	QM	F_0
Regressão	1	SQR_{reg}	SQR_{reg}	$\frac{SQR_{reg}}{SQE/(n-2)}$
Erro	$n-2$	SQE	$SQE/(n-2)$	
Total	$n-1$	SQT		

Regressão Múltipla

Regressão Linear Multivariada

- Relembre a equação para a regressão linear univariada:

$$\hat{y} = \beta_0 + \beta_1 x$$

- A equação de regressão linear multivariada é a seguinte

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

- Assume-se que a variável respostas é uma combinação linear das variáveis preditoras

Alertas sobre Regressão Linear

- Verifique se seus dados são realmente lineares
- Tenha certeza que o tamanho da amostra é suficiente
- Não use um modelo de regressão para predizer dados fora das dimensões dos dados utilizados fora do modelo
- Resultados são muitos sensíveis na presença de outliers
- Regressão Múltiplica: cheque se suas variáveis preditoras são independentes
- Regressão mostra correção, mas não necessariamente casualidade

Mais dicas

- Faça scatter plots dos dados e veja se existe linearidade
- Grandes amostras são melhores
- A melhor forma de tratar outliers depende dos objetivos da sua análise (tente plotar os dados)
- Elimine os outliers que você tenha certeza que indicam erro nos dados
- Tente criar modelos com e sem outliers

Modelos não-lineares

Como avaliar o modelo?

- Exemplo: Perde $ devido a lista de preço errado
  - Muito baixo ??
  - Muito alto ??

  O quanto eu estou perdendo comparado com a perfeição?
  Predição perfeita: $Loss = 0$
  Minhas predições: $Loss =$??

Medindo a perda

Função de perda

$$L(y, f_w(x))$$

- $f_w(x) = \hat{f}(x)$ – função de perda
- Exemplos:
  - Erro absoluto: $L(y, f_w(x)) = |y - f_w(x)|$
  - Erro quadrado: $L(y, f_w(x)) = (y - f_w(x))^2$

"Lembre-se que todos os modelos estão errados; a questão prática é o quanto errado eles estão para não ser úteis"
George Box, 1987

Computando o Erro

- Definindo a função de erro $L(y, f_w(x))$
  - erro quadrático, erro absoluto, ...
- Erro no conjunto de treinamento
  - erro médio no conjunto de treinamento
  - $\frac{1}{N} \sum_{i=1}^{N} L(y, f_w(x_i))$
  - ajuste feito usando os dados de treino
- Calculando o erro no treinamento vs a complexidade do modelo

Boas medidas de performance

- O erro no treinamento é uma boa medida para a performance do modelo?
- Existe alguma coisa particularmente ruim sobre fazer o quadrado de x_t?
- O treinamento pode ser super otimista, diminuindo o erro.
  - os dados podem ser super ajustados no modelo
  - Pequenas quantidades de dados é inversamente proporcional a bons preditores.

Generalizando a definição de erro

- Estimando a perca sobre todos os possíveis exemplos
- Não é possível computar o erro para todas as possibilidades de entradas
- Deve-se fazer uma aproximação olhando para exemplos que não estão nos conjuntos de dados.
- Dividir em
  - Treino
  - Teste

Computando o erro

Erro no teste = a perca média no conjunto de teste

$$= \frac{1}{N_{teste}} \sum_{i \quad no \quad teste} L(y_i, f_w(x_i))$$

Divida o conjunto de dados em treino e teste

Viês e variância

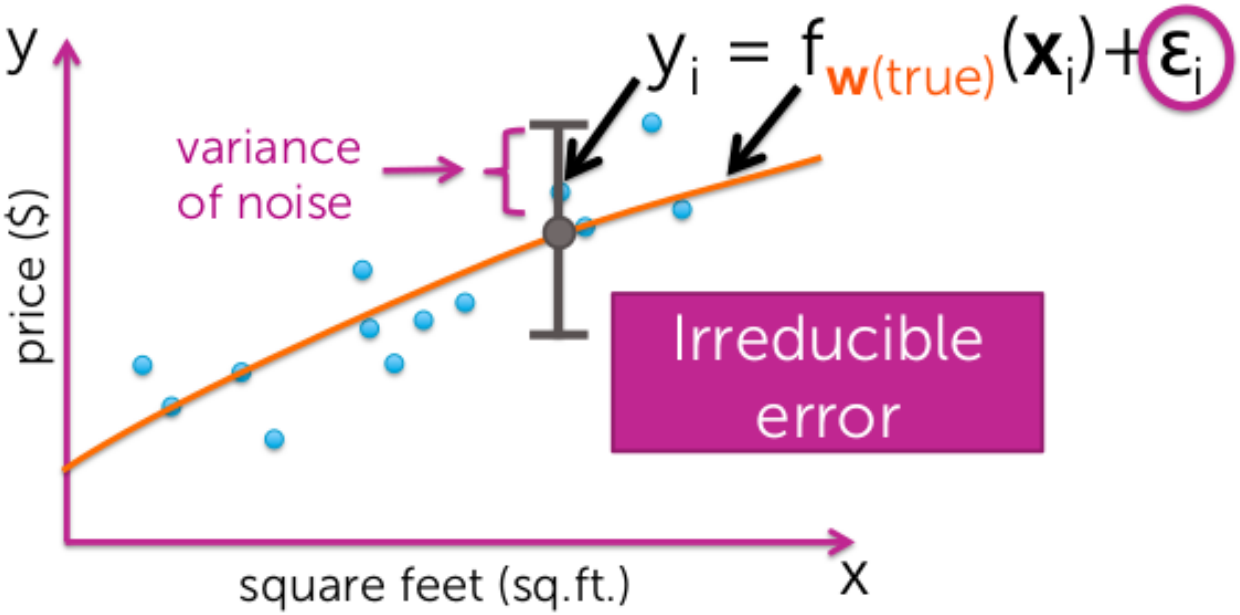
bias-variance tradeoff

- Existem três formas de erros que se deve considerar em modelos de regressão
  1. Ruído
  2. Viês
  3. Variancia

Ruídos

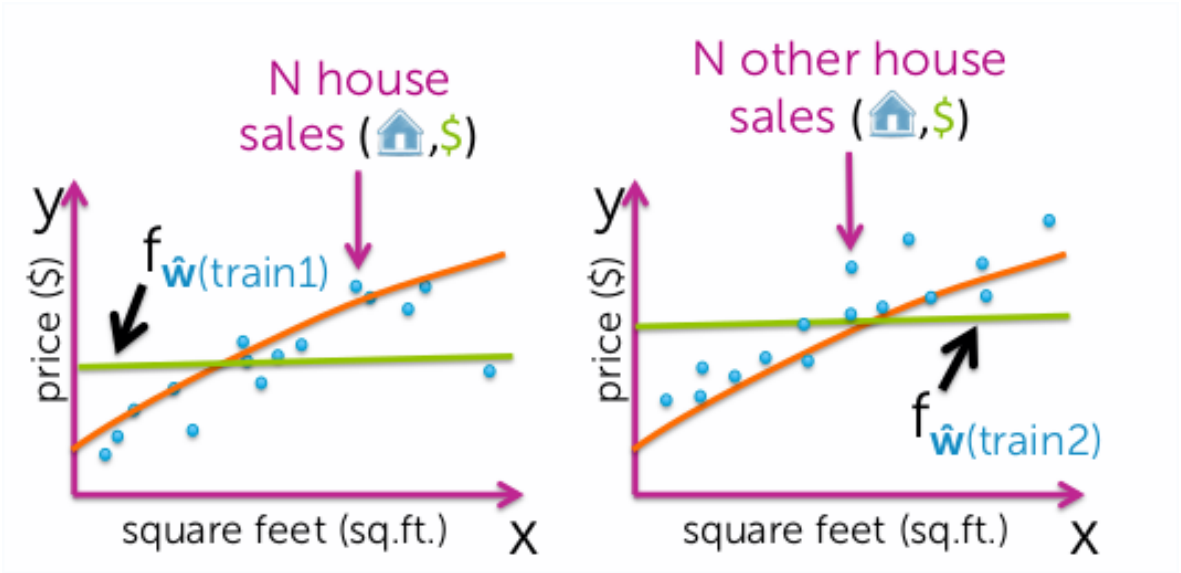
- Variância no ruído indica erro irredutível
- Alguns modelos possuem a seguinte função de predição
  - $y_i = f_w(x_) + \epsilon$

Ruídos


y
price ($)
x
square feet (sq.ft.)
$y_i = f_{w(true)}(x_i) + \varepsilon_i$
variance of noise
Irreducible error

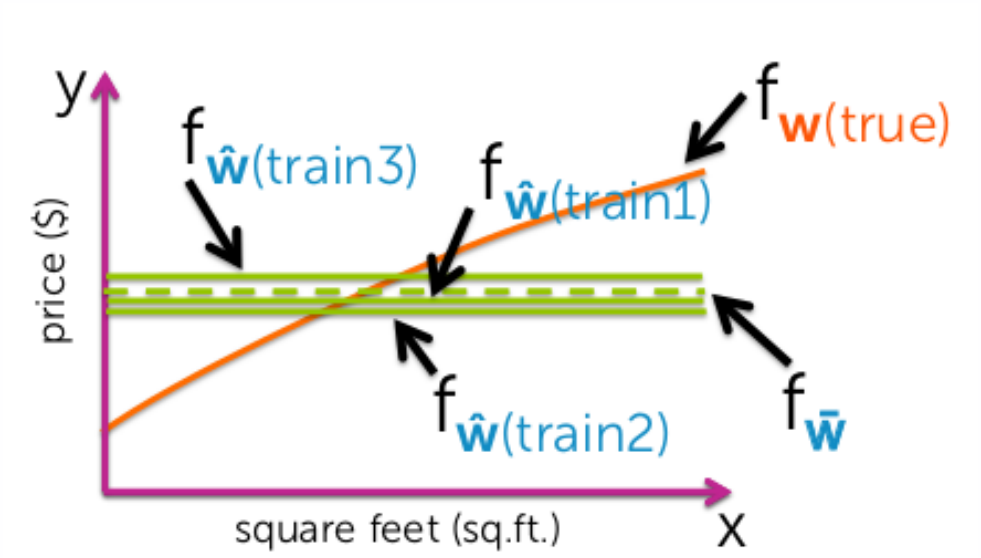
Contribuição do viés

Suponha que ajustamos uma função constante



Contribuição do viés

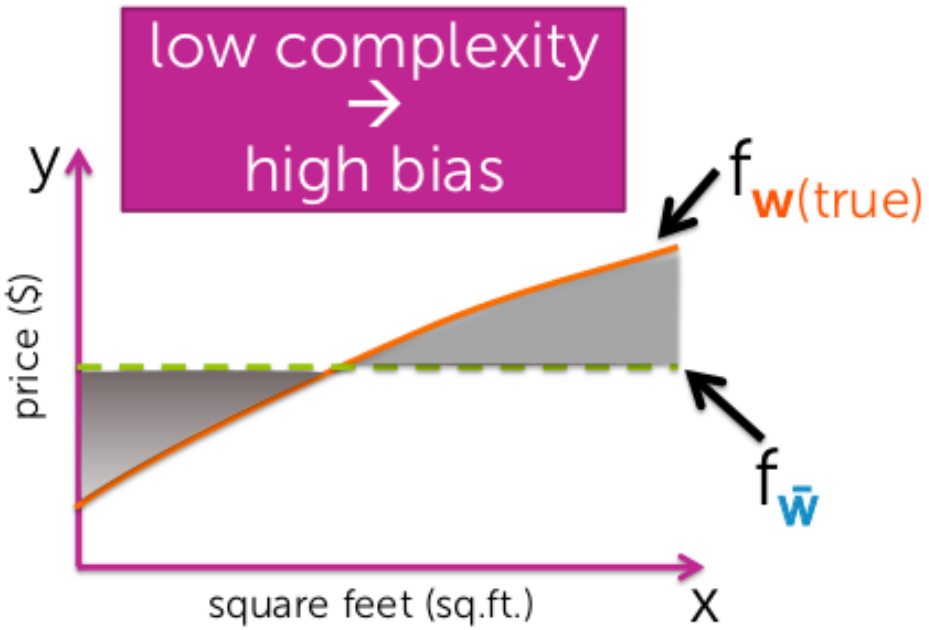
Sobre todos os conjuntos de treinos de tamanho N possíveis, o que é esperado do ajuste?



Contribuição do viés

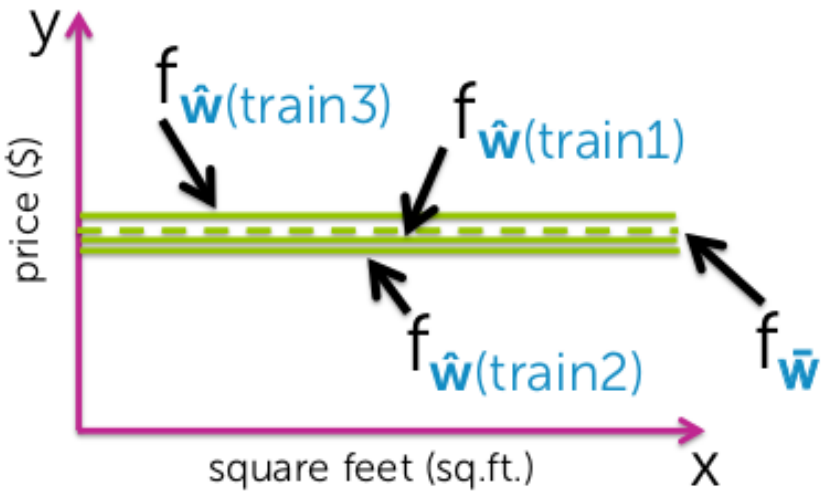
$Bias(x) = f_{w(true)}(x) - f_w(x)$

Nosso modelo é flexivível suficiente para capturar $f_{w(true)}$? Se não, existem erros na predição.



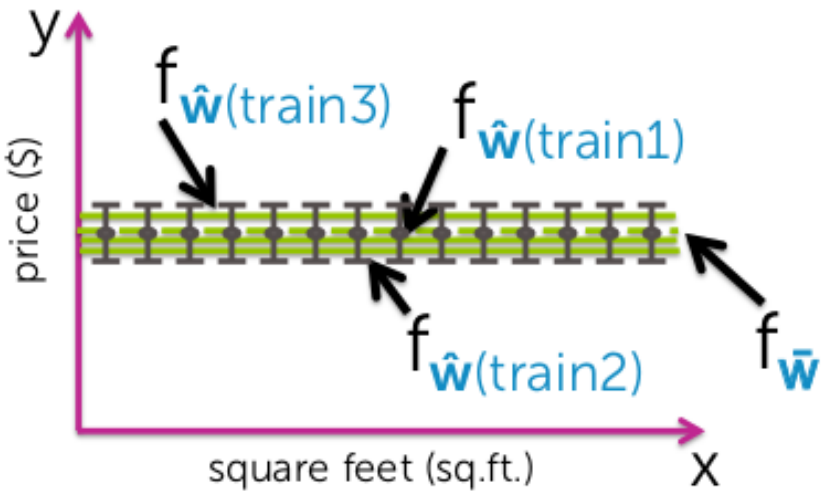
Contribuição da variância

O quanto o ajuste do modelo específico varia do modelo esperado?



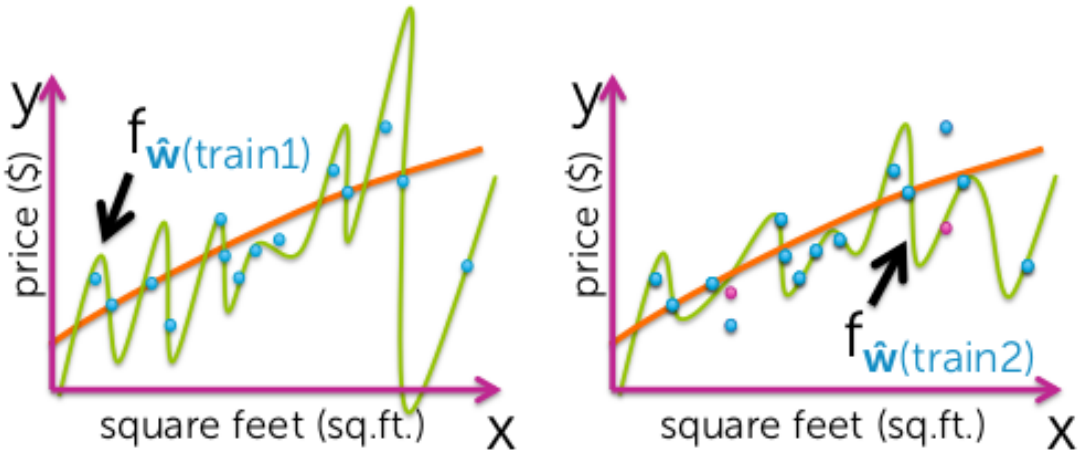
Contribuição da variância

O quanto o ajuste do modelo específico varia do modelo esperado?



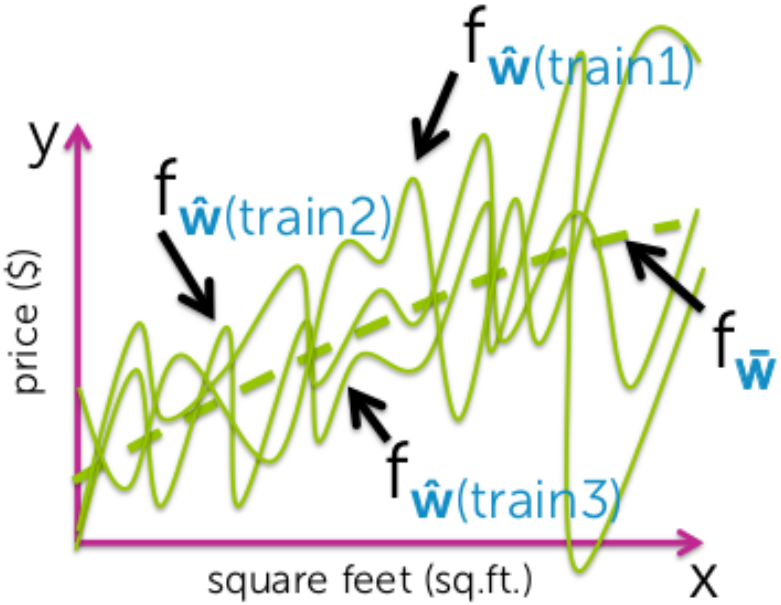
Variância em modelos complexos

Suponha que ajustamos um polinômio de alta ordem



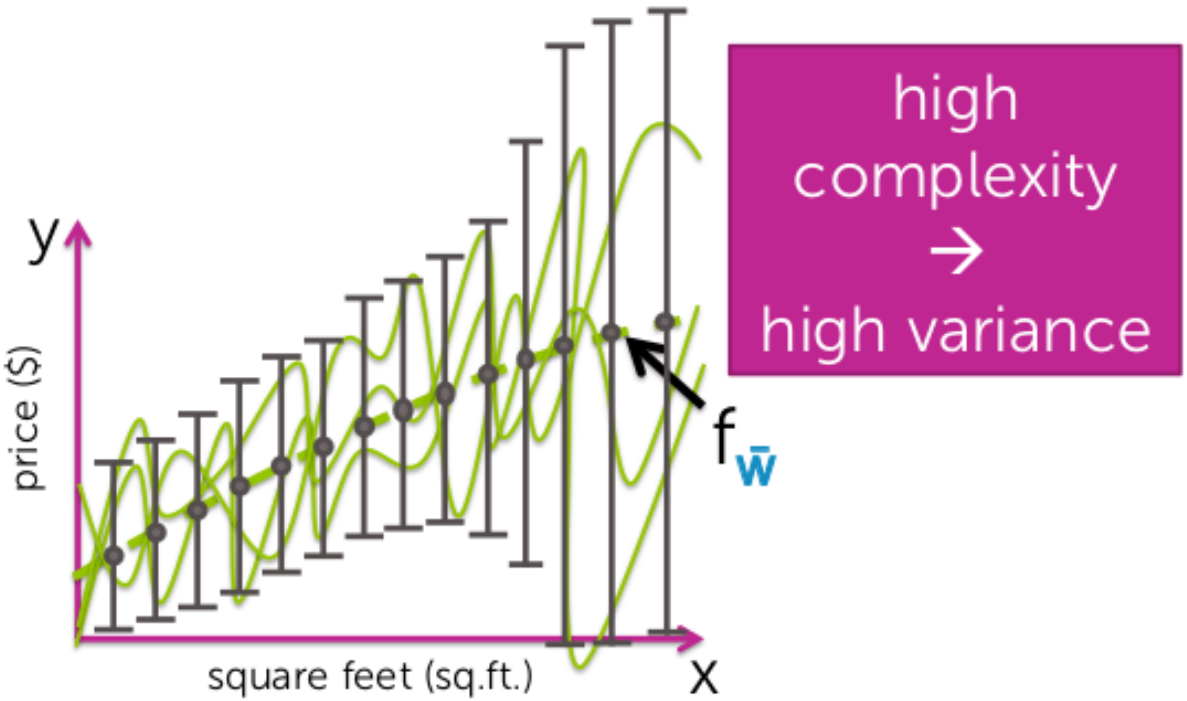
Variância em modelos complexos

Suponha que ajustamos um polinômio de alta ordem

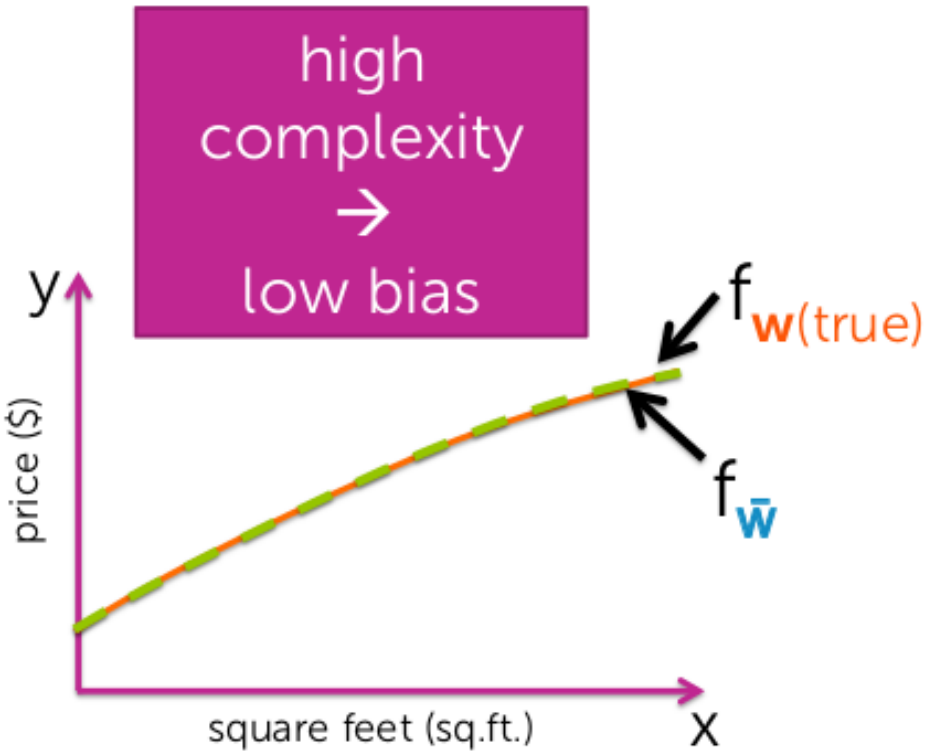


Variância em modelos complexos

Suponha que ajustamos um polinômio de alta ordem



Viés em modelos complexos



Erro esperado

Erro esperado do preditor para um exemplo x_t

$$= \epsilon + MSE[f_w(x_t)] = \epsilon + [bias(f_w(x_t)]^2 + var[f_w(x_t)]$$

- Veja as três fontes de erro na equação!

O fluxo de Regressão em tarefas de aprendizado de máquinas

- Seleção do modelo
  Frequentemente, é necessário escolher os parâmetros λ que controlam a complexidade do modelo
- Custo do modelo
  Uma vez selecionado o modelo, calcule o custo

Implementação Hipotética

1. Seleção de modelo
   Para cada modelo com complexidade λ
   1. Estime os parâmetros w no conjunto de dados de treino
   2. Calcule a performance do modelo no conjunto de teste
   3. Escolha o conjunto λ^* com o menor erro no conjunto de teste
2. Custo do modelo
   Compute o erro no conjunto de teste para aproximar o erro generalizado

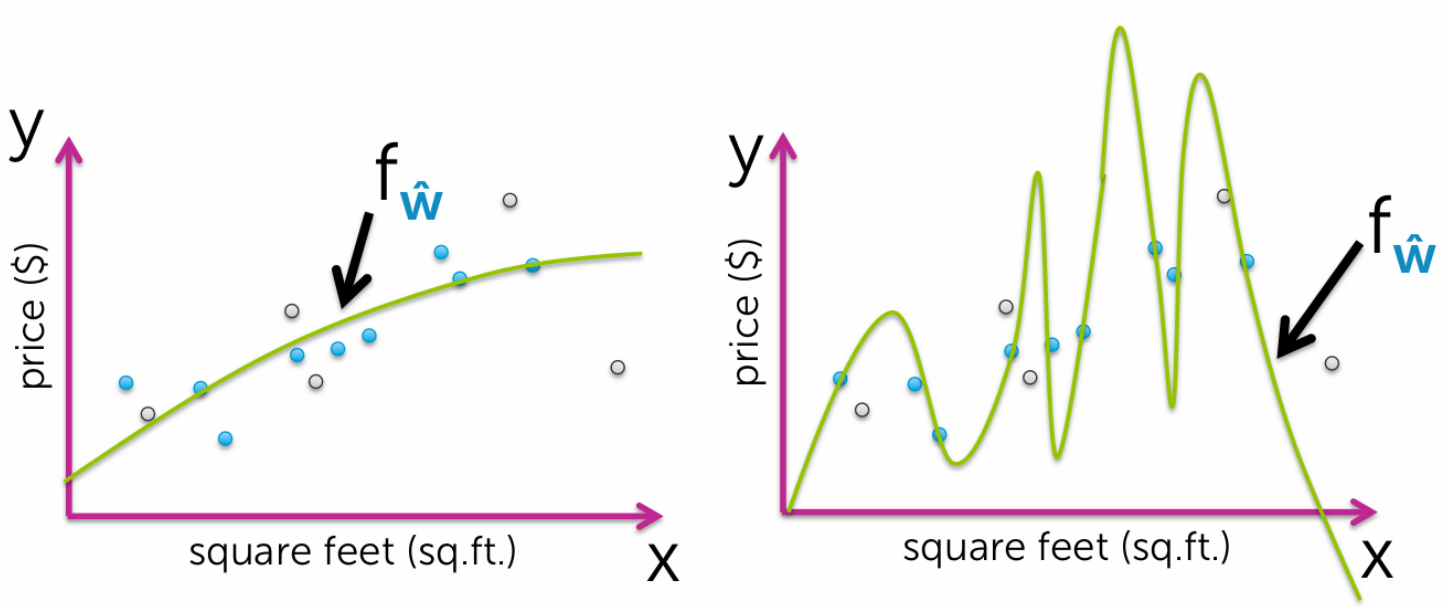
Implementação Hipotética

- Se o conjunto de dados do teste não é representativo, então w_λ será tipicamente pior do que o indicado.
  - Solução: Crie dois conjuntos de teste!
    1. Selecione os parâmetros λ tal que minimize w_λ no conjunto de validação
    2. Aproxime o erro generalizado utilizando o conjunto de teste.

Overfitting em regressores polinomiais

Flexibilidade em polinomiais em alta ordem

$$y_i = w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p + \epsilon_i$$



Overfitting em vários atributos

Frequentemente, overfitting é associado a valores altos de parâmetros w

- Poucas observações – rapidamente a complexidade do modelo aumenta, logo aumenta o overfitting
- Várias observações – difícil de dar overfitting

Os dados devem incluir exemplos representativos de todos os conjuntos. Isso é difícil!

Formato do custo total

O que queremos?

1. O quanto nossa função se ajusta nos dados
2. Magnitudo dos coeficientes

Custo total =
medida de ajuste + medida de magnitude dos coeficientes

Medida de ajuste dos dados

- Erro quadrádito

$$MSE = \sum_{i}^{N} (y_i - h(x_i)^T w)^2$$

$$= \sum_{i}^{N} (y_i - \hat{y}_i(w))^2$$

Medida de magnitude dos coeficientes

Como indicar se os coeficientes estão altos?

- soma
- soma dos valores absolutos (norma l_1)
- Soma dos quadrados (norma l_2)

Custo total

Custo total =
medida de ajuste + medida de magnitude dos coeficientes

- medida de ajuste – MSE
- magnitude dos coeficientes – norma l_2 $(||w||_2^2)$

Considerando os objetivos

Queremos encontrar o mínimo valor de w na seguinte função de custo:

$$MSE(w) + \lambda||w||_2^2$$

Ridge Regression

$$MSE(w) + \lambda ||w||_2^2$$

- Valor alto de λ significa alto bias, e baixa variância
- Valor baixo de λ significa baixo bias, e alta variância

Revisitando o ajuste de modelos polinomiais

Ridge Regression

Notação matricial

$$MSE(w) + \lambda ||w||_2^2$$
$$= (y - Hw)^T (y - Hw) + \lambda w^T w$$

Como otimizar isso?

Gradiente da função de custo do Ridge Regression

$$\begin{aligned}\nabla MSE(w) + \lambda||w||_2^2 &= \nabla[(y - Hw)^T(y - Hw) + \lambda w^T w] \\ &= \nabla[(y - Hw)^T(y - Hw)] + \nabla[\lambda w^T w] \\ &= 2H^T(y - Hw) + \lambda 2w\end{aligned}$$

Equação de atualização

$$w_j^{(t+1)} \leftarrow w_j^{(t)} - \eta[-2\sum_{i=1}^{n}(h(x_i)(y_i - \hat{y}_i(w^{(t)}))] + 2\lambda w_j$$

Algoritmo do Regressor de Ridge

1. inicie $w^{(1)} \leftarrow 0$, para $t = 1$
2. enquanto $||\nabla MSE(w^{(t)}|| > \epsilon$
    1. para $j \leftarrow$ até D (número de dados)
        1. $partial[j] \leftarrow -1 \sum_{i=1} N(Y_i - \hat{y}(w^{(t)}))$
        2. $w^{(t+1)} \leftarrow w_j^{(t)} - \eta partial[j]$
        3. $t \leftarrow t + 1$

Como escolher λ

Escolha o valor do λ no conjunto de validação

- Use a validação cruzada
  - Normalmente para k $= 5$ ou k $= 10$
- A melhor aproximação ocorre quando se usa o $leve - one - out$

Lasso Regression

A tarefa de extração de características

Por que queremos fazer a seleção de características?

- Se o tamanho do vetor w é 100B, cada predição é cara!
- Se o vetor w for esparso, a computação depende apenas dos valores não-zeros

Quais características são relevantes para a predição

Encontre o melhor modelo

O ideal seria criar um modelo para cada subconjunto de características.
Mas isso é possível?

Encontre o melhor modelo

É inviável procurar o melhor modelo para os 2^M subconjuntos de features.

$$2^{30} = 1073741824$$

$$2^{1000} \approx 1x10^{301}$$

Podemos usar uma solução gulosa!

Forward stepwise algorithm

- Escolha uma característica
- Heurística gulosa
  - Comece com o conjunto de características vazias $F_0 = 0$ $(y_i = w_0 + \epsilon)$
  - Ajuste o modelo usando as características do conjunto F_t para obter os coeficientes $w^{(t)}$
  - Selecione as melhores características $h_j(x)$
    - $h_j(x)$ resultando no menor erro no conjunto de treino quando se aprende com $F_t + \{h_j(x)\}$ características
  - Defina $F_{t+1} \leftarrow F_t + \{h_j(x)\}$
  - Continue...

Qual a complexidade desse algoritmo?

Quando parar

- Quando o erro no treinamento é baixo o suficiente?
- Quando o erro no teste é baixo o suficiente?

Quando parar

- Quando o erro no treinamento é baixo o suficiente?
- Quando o erro no teste é baixo o suficiente?

Use validação cruzada!

Ridge Regression

Custo total =
medida de ajuste + λ medida de magnitude dos coeficientes

Isso encoraja pesos pequenos, mas não exatamente iguais a 0

Modelos esparsos

- Eficiência
  - Se o número de características for alto?
- Interpretabilidade
  - Quais características são relevantes para predição?
- Interpretabilidade

Usando regularização para seleção de características

- Em vez de buscar sobre um conjunto discreto de soluções, podemos usar regularização?
  - Comece com o modelo completo
  - Reduza alguns coeficientes para zero
  - Coeficientes não-zero indicam boas características

Lasso Regression

Custo total =
medida de ajuste + λ medida de magnitude dos coeficientes

- Medida de ajuste é o MSE(w)
- Já a magnitude dos coeficientes é calculado com a norma l_1

$$||w|| = |w_0| + ... + |w_D|$$

Como otimizar o Lasso Regressor

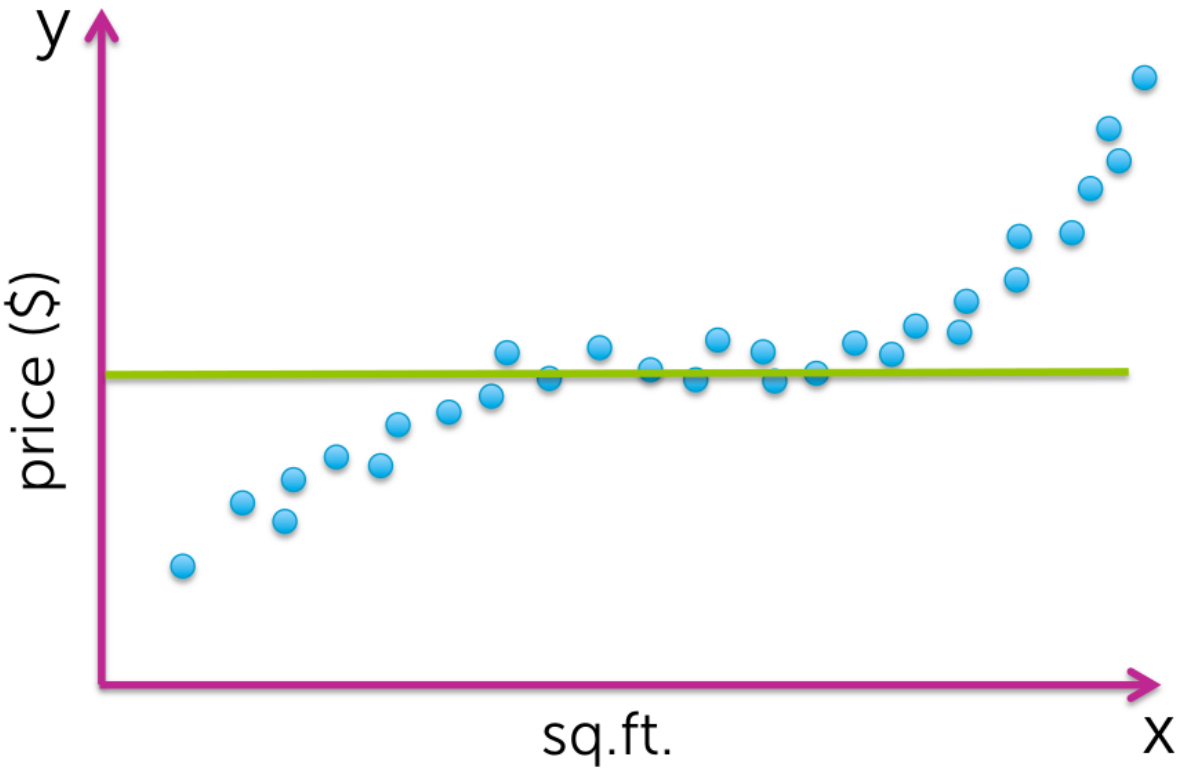
- Utiliza-se o método da descida coordenada
- Otimiza a função de custo $g()$ para cada coordenada
  - Inicia $w \leftarrow 0$ (ou número pequeno)
  - enquanto não converge
    - escolha uma coordenada j
    - $w_j \leftarrow min_w g(w_0, w_1, ..., w, w_{j+1}, ..., w_D)$
- Como escolhemos a próxima coordenada
  - escolhido aleatoriamente
- Não necessita de passos para otimizar
- Técnica útil para vários problemas
  - Converge para o ótimo em alguns casos
  - Converge para a função objetivo do Lasso

Escolhendo o valor de λ

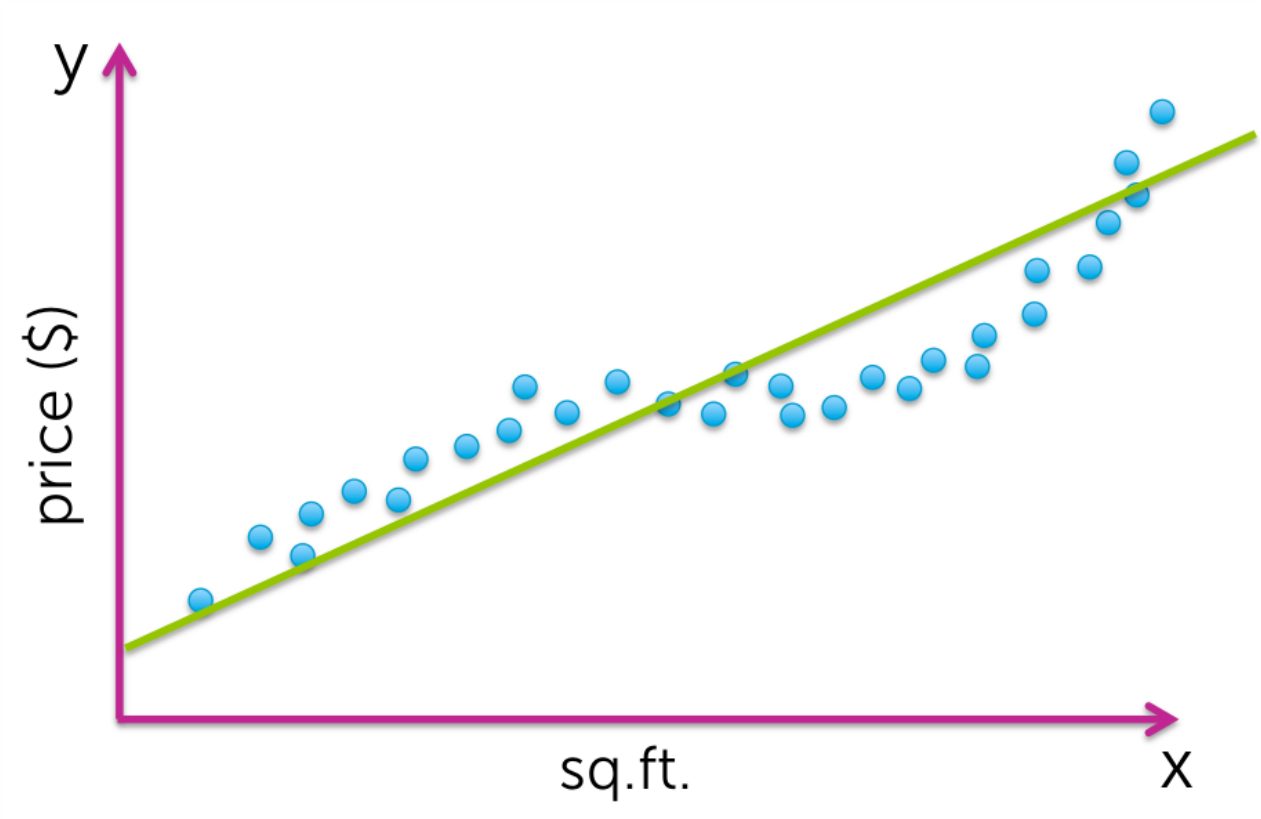
- Validação cruzada para escolher λ que prevê o melhor modelo preditivo
- Tende-se a escolher valor λ baixo, consequentemente menos valores esparsos.

Modelos Não-Paramétricos

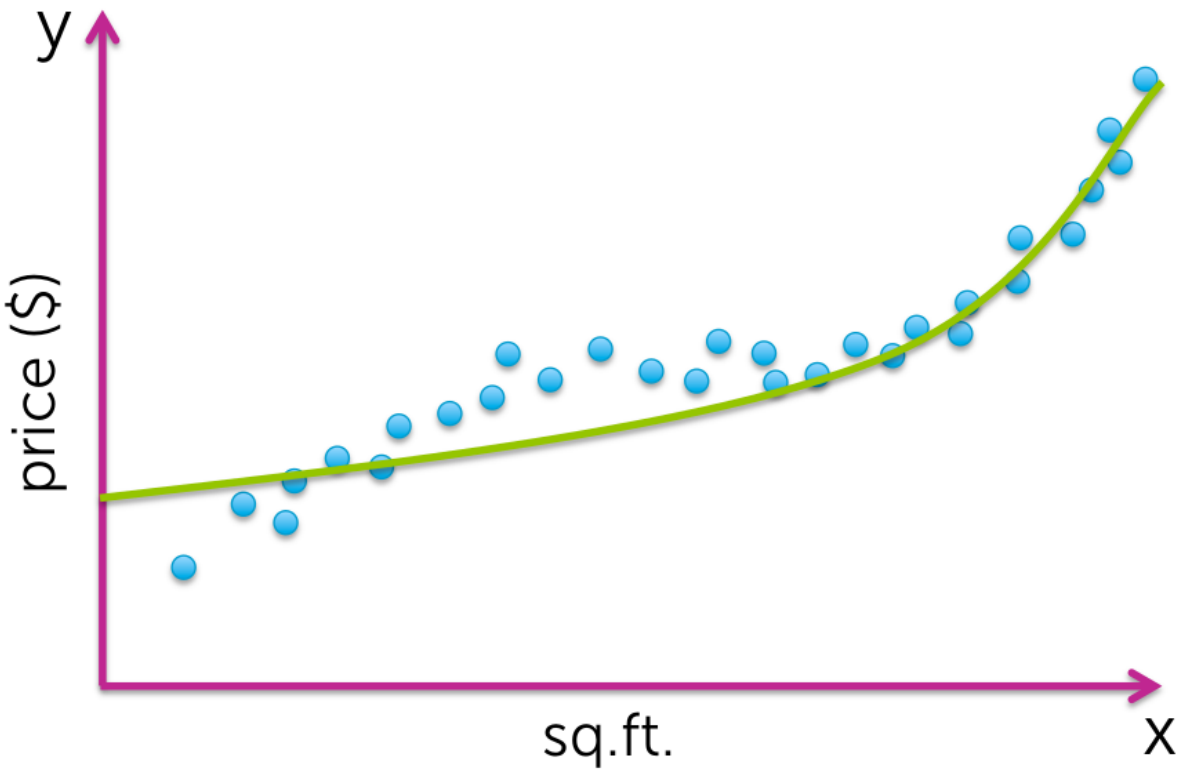
Modelos paramétricos



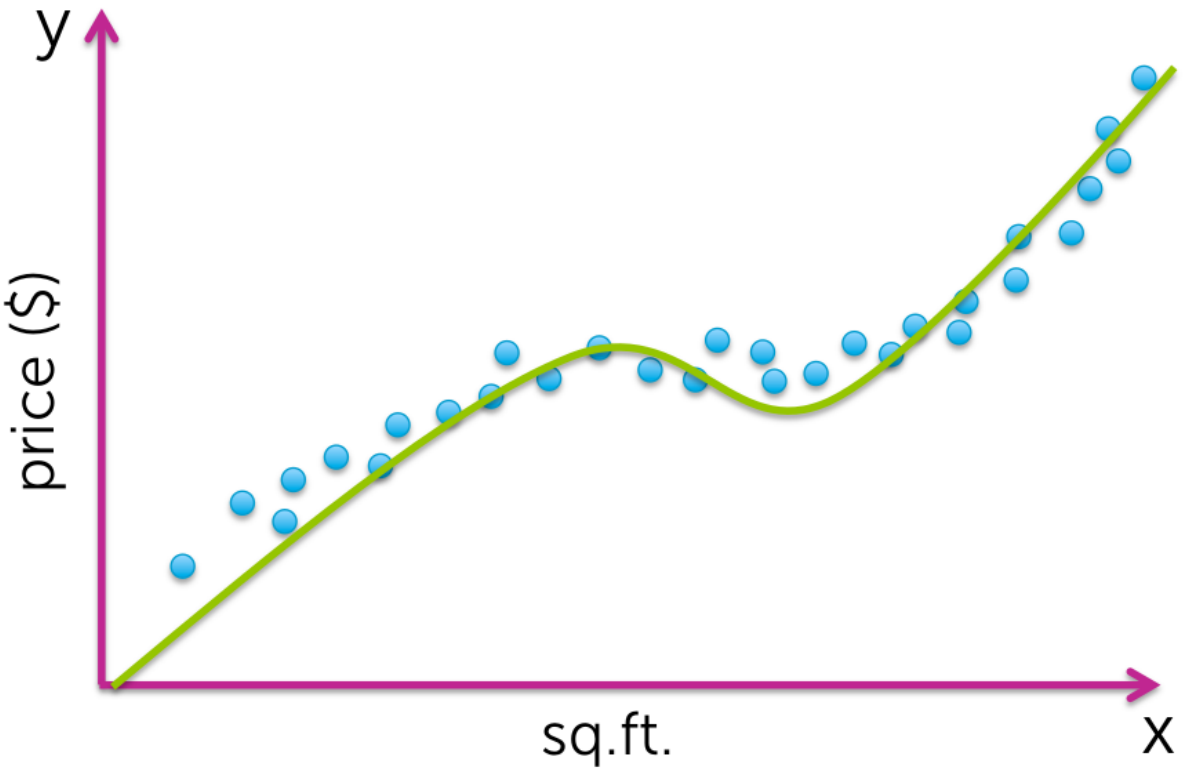
Modelos paramétricos



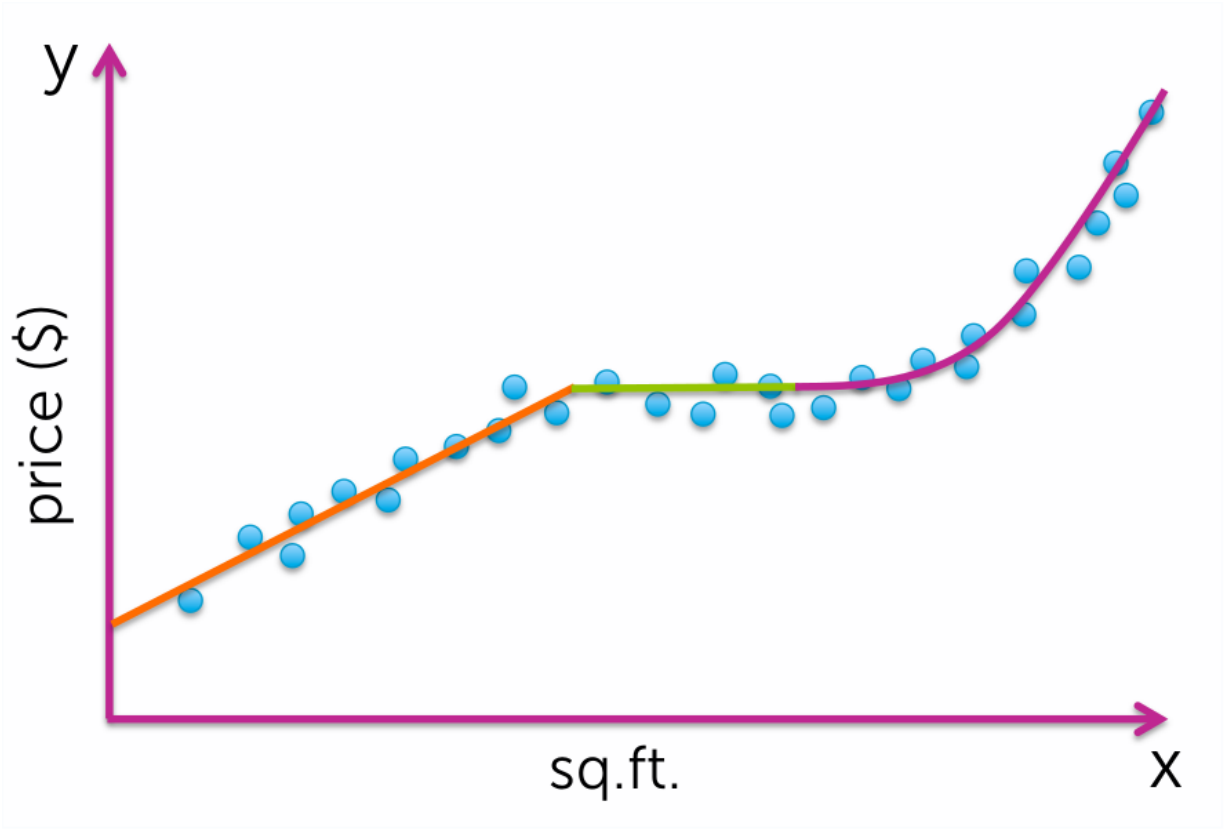
Modelos paramétricos



Modelos paramétricos



$f(x)$ não é necessariamente polinomial



Quais alternativas nos temos?

E se:

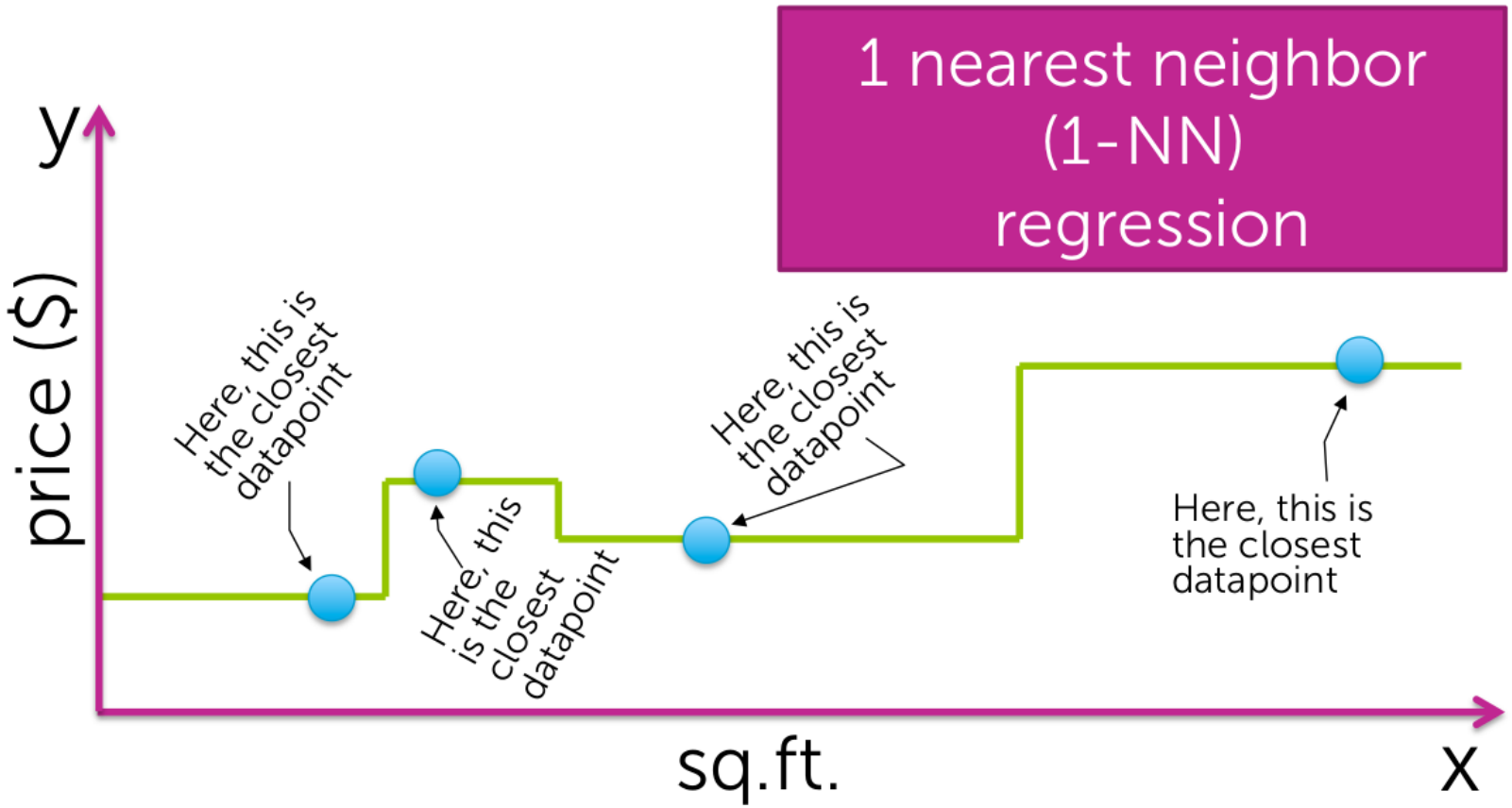
- Queremos flexibilizar a função $f(x)$ para ter estrutura local?
- Não precisamos inferir quebras no conjunto de dados

Quai a opção simples que temos?

- Assumimos que temos grande conjunto de dados

1-Vizinho mais próximo

prediz valor = o y_i mais próximo
Parece o que é feito em estimação "manual"



1-NN regression

Dado um query point x_q

1. Encontre o ponto x_i mais próximo
   - $x_{NN} \leftarrow min_i dist(x_i, x_q)$
2. Predict
   - $\hat{y}_q \leftarrow y_{NN}$

Medidas de distância

- Em 1-D, apenas a distância Euclidiana

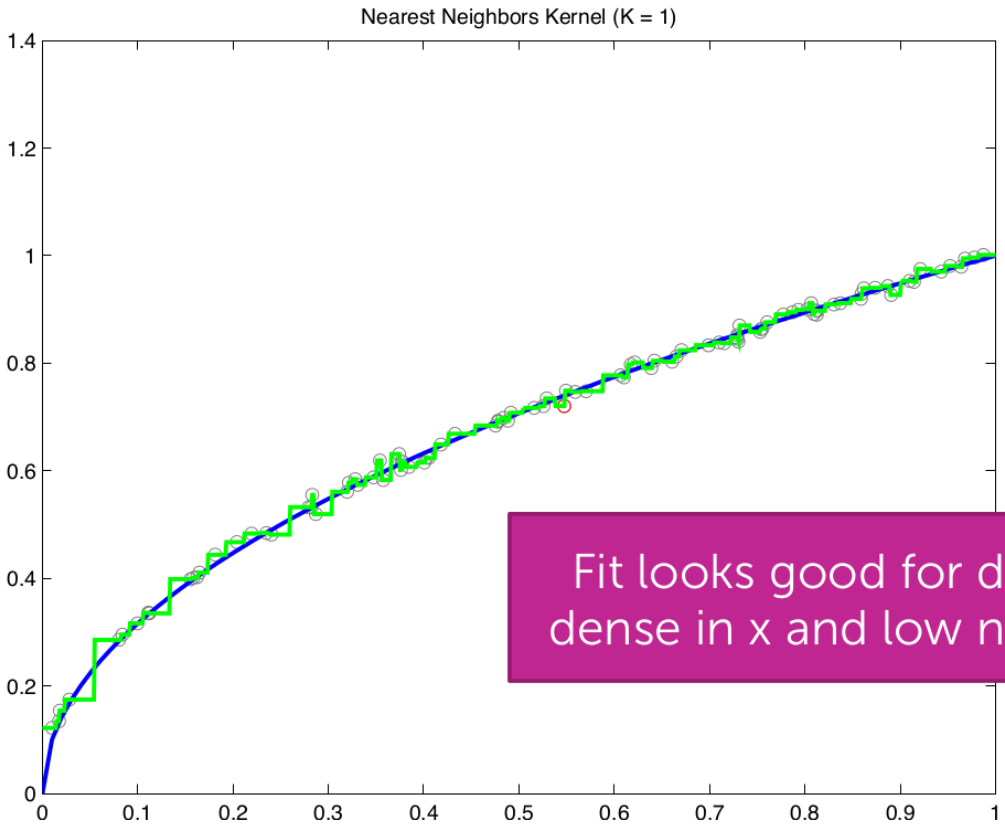
$$dist(x_i, x_q) = |x_i - x_q|$$

- Em múltiplas dimensões
  - podemos definir várias funções de distância

Distância ponderada

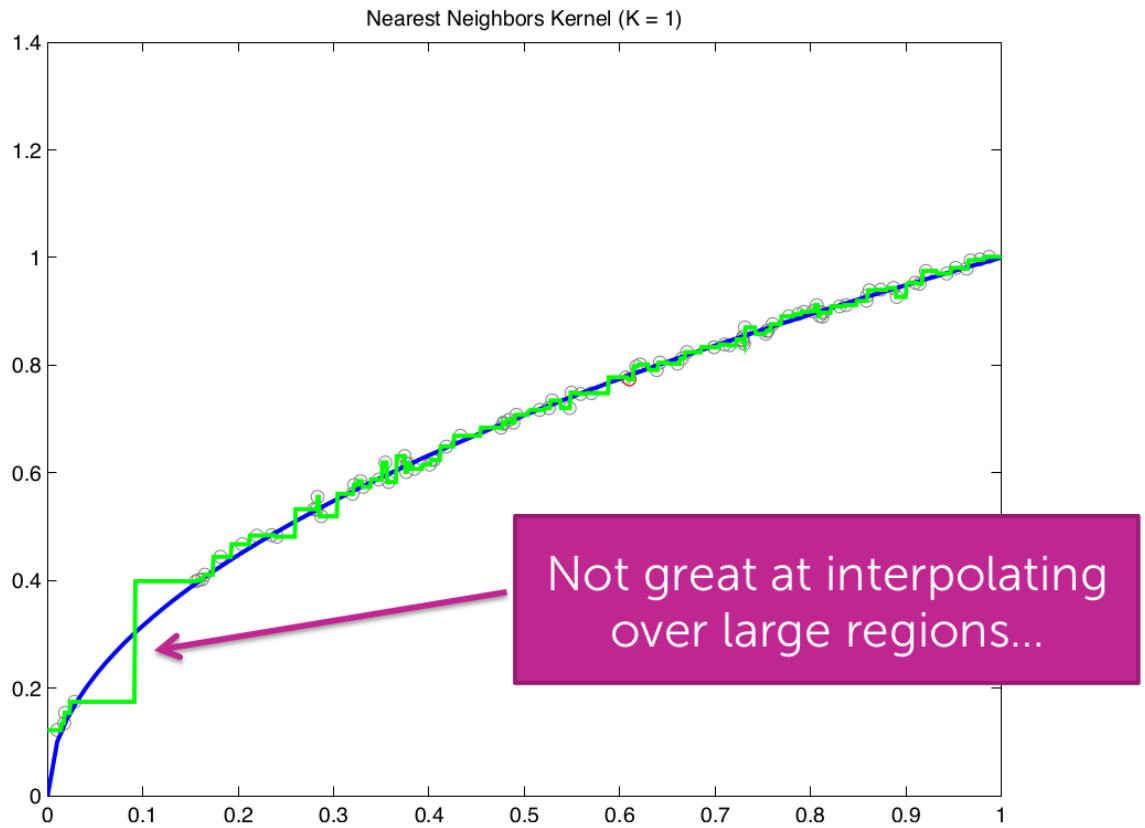
$$dist(x_j, x_q) = \sqrt{a_1(x_j[1] - x_q[1])^2 + ... + a_d(x_j[d] - x_q[d])^2}$$

1-NN na prática



Fit looks good for data dense in x and low noise

1-NN na prática



Dado uma query x_q

1. Encontra os K vizinhos mais próximos
   - $(x_{nn1}, x_{nn2}, ..., x_{nnk})$ tal que para qualquer x_i não está no conjunto dos k vizinhos
2. Predição

$$\hat{y}_q = \frac{1}{k}(y_{nn1}, y_{nn2}, \ldots, y_{nnk})$$
$$= \frac{1}{k}\sum_{j}^{k} y_{nnj}$$

Weighted K-NN

Podemos dar pesos para exemplos mais próximos na lista dos K vizinhos mais próximos

Predict:

$$\hat{y} = \frac{(c_{qnn1}y_{nn1} + c_{qnn2}y_{nn2} + \ldots = c_{qnnk}x_{nnk})}{\sum_j^k c_{qnnj}}$$

Como definir os pesos

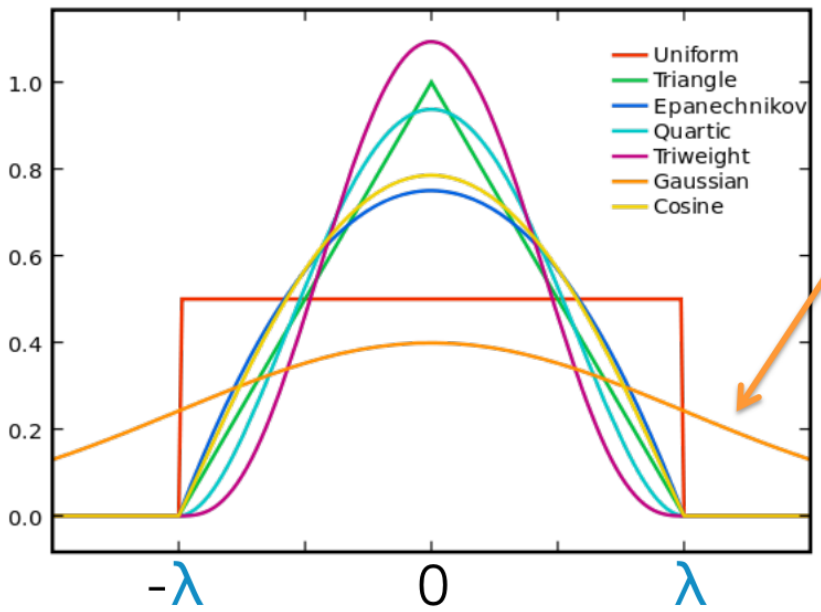
Queremos dar pesos pequenos para c_{qNNj} quando a distância é grande e c_{qNNj} grande quando a distância é pequena

$$c_{qnnj} = \frac{1}{dist(x_j, x_q)}$$

Kernel

Define

$$c_{qnnj} = kernel_{\lambda}(|x_{nnj} - x_q|)$$



Gaussian kernel:

$Kernel_{\lambda}(|x_i - x_q|) = \exp(-(x_i - x_q)^2/\lambda)$

Note: never exactly 0!

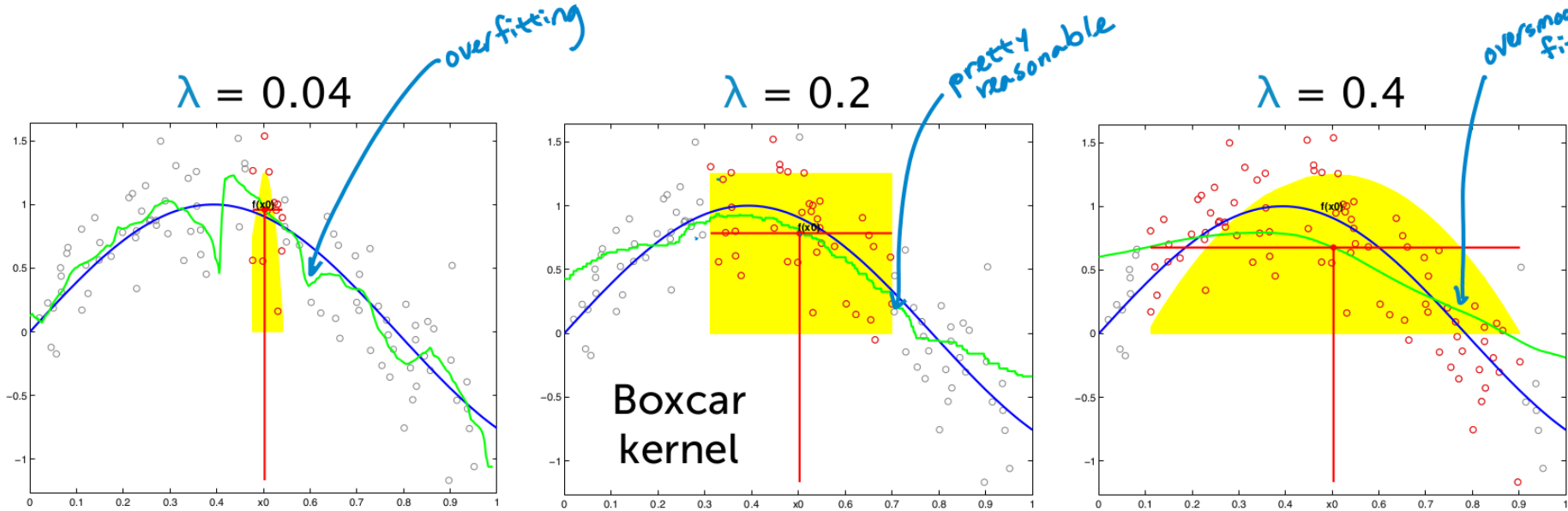
Kernel Regression

Em vez de atribuir peso para cada vizinho, é atribuido peso a cada ponto

Predição:

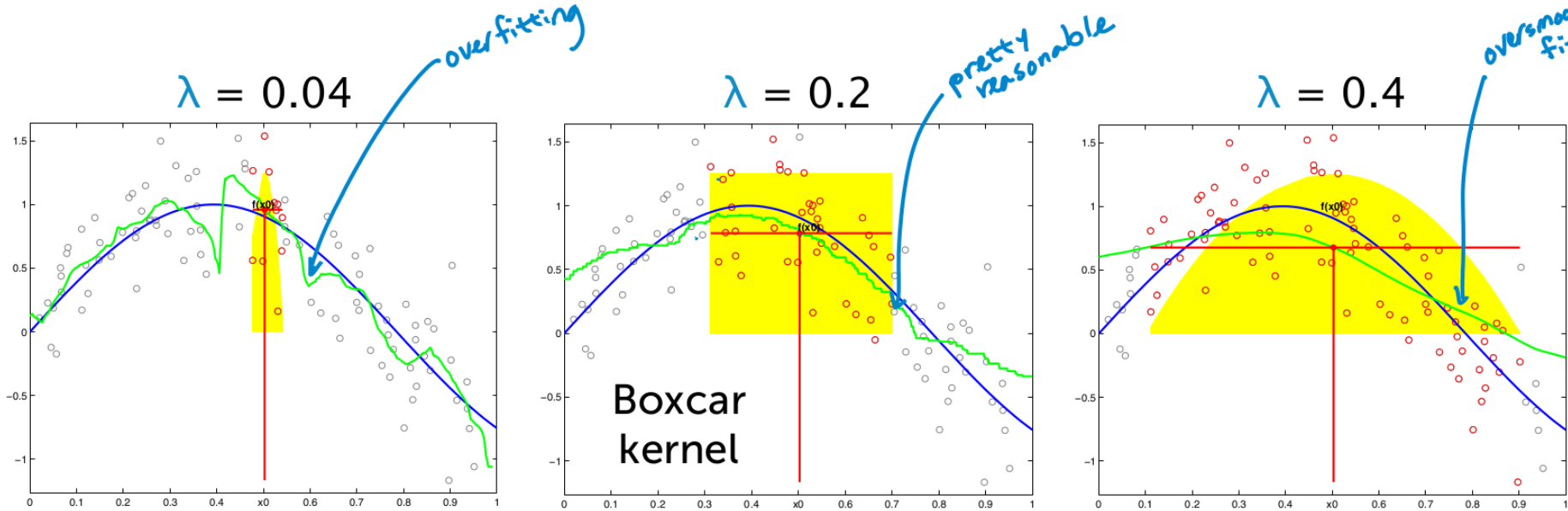
$$\hat{y}_q = \frac{\sum_i^N c_{qi} y_i}{\sum_i^N c_{qi}} = \frac{\sum_i^N kernel_\lambda(dist(x_i, x_q))y_i}{\sum_i^N kernel_\lambda(dist(x_i, x_q))}$$

Escolha do parâmetro do kernel



Como escolher o λ??

Escolha do parâmetro do kernel



Como escolher o λ??

Validação Cruzada

Knn e Kernel Regression

KNN e Kernel Regression são exemplos de regressores não parametricos

Objetivo geral dos regressores não parametricos

- Flexibilidade
- Criar suposições sobre $f(x)$
- Complexidade pode aumentar com o numero de exemplos

Knn e Kernel Regression

KNN e Kernel Regression trabalham melhores com grande quantidade de exemplos

- Quanto maior o número de dimensões maior o número de pontos
- espera-se que $N = O(exp(d))$ exemplos

- Dado um ponto x_q como definir a distância?
- $O(N)$ para 1-NN
- $O(N \log k)$ para o k-NN

Árvores regressoras

Relembrando

- Função objetivo geral

$$Obj(w) = L(w) + \Omega(w)$$

onde $L(w)$ é o erro no treinamento, medindo o quanto bem o modelo ajusta os dados, $\Omega(w)$ mede a complexidade do modelo

- Ajustando o modelo – $L = \sum_i^N l(y_i, \hat{y}_i)$
  - Erro quadrático: $l(y_i, \hat{y}_i) = (y_i - \hat{y})^2$
  - Regressão Logística:
    $l(y_i, \hat{y}_i) = y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i) \ln(1 + e^{\hat{y}_i})$
- Regularização: O quanto complicado o modelo é?
  - norma l_2: $\lambda ||w||_2^2$
  - norma l_1: $\lambda ||w||_1$

Colocando todo o conhecimento de contexto

- Ridge regression: $\sum_i^N (y_i - w^T x_i)^2 + \lambda ||w||_2^2$
  - Modelo linear, erro quadrático, regularização l2
- Lasso: $\sum_i^N (y_i - w^T x_i)^2 + \lambda ||w||_1$
  - Modelo linear, erro quadrático, regularização l2
- Regressão Logística:
  $\sum_i^N [y_i \ln (1 + e^{-w^T x_i}) + (1 - y_i) \ln (1 + e^{w^T x_i})] + \lambda ||w||^2$
  - Modelo linear, regressão logística, regularização l2
- A separação conceitual entre modelo, parâmetros, objetivos também dão benefícios para projetar o algoritmo
  - Como otimizar essas funções objetivos

Relembrando

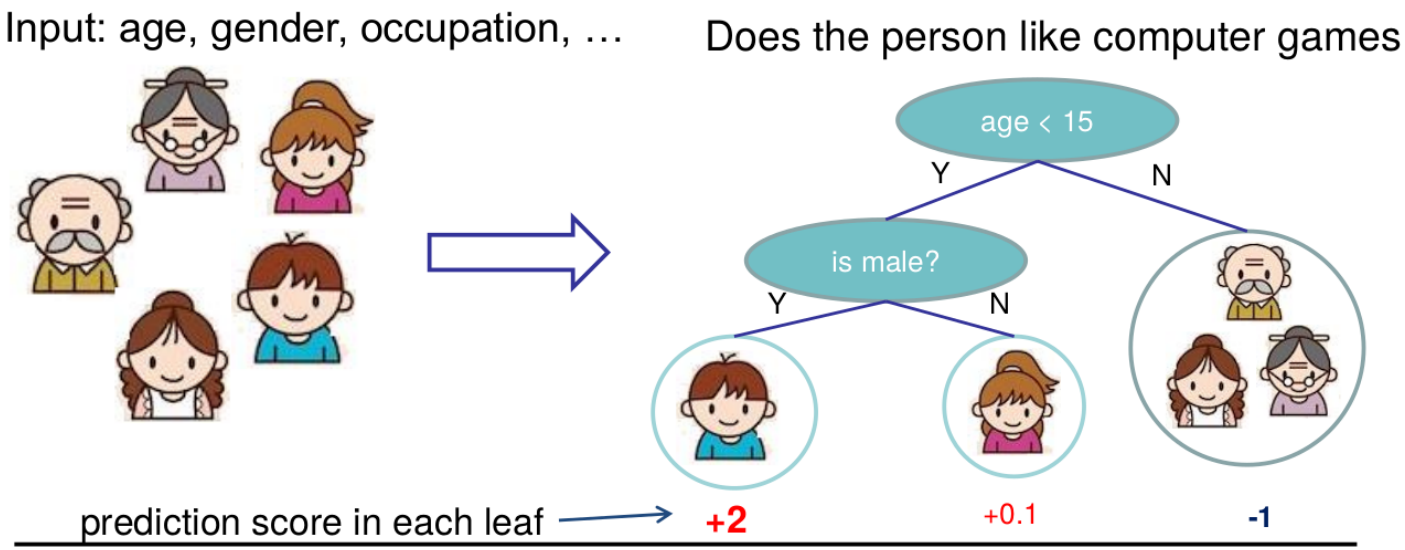
- Função objetivo geral

$$Obj(w) = L(w) + \Omega(w)$$

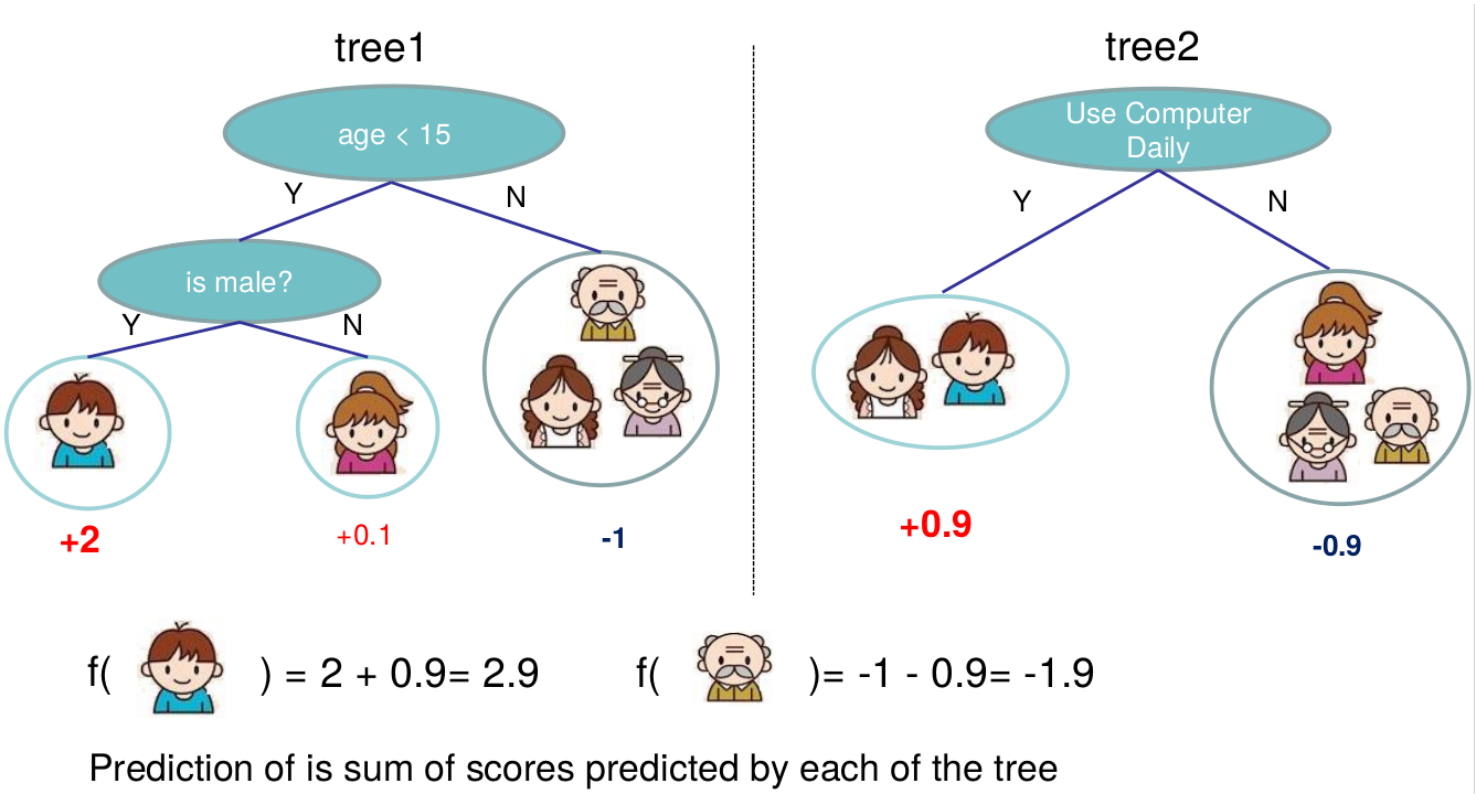
- Otimizando L encorajamos bons modelos preditivos
  - diminui o bias
- Otimizando Ω encoraja modelos simples
  - diminui a variância

Árvore de Regressão

- Árvore de regressão contém a mesma estrutura de uma árvore de decisão
- Contem um socore em cada nó folha da árvore



Comitê de árvores Regressoras



Ensembles methods

- Técnica muito utilizada, procure por GBM, random forest,...
  - quase metade das competições de data mining são vencidas usando alguma variante de métodos ensemble de árvores
- Invariante em escala de entrada, e você não precisa se preocupar com normalização de características
- Aprende interações de alta ordem em características
- Pode ser escalável, e são usadas em aplicações práticas

Comitê de Árvores Regressoras

- Modelo: assumindo que nós tenhamos K árvores

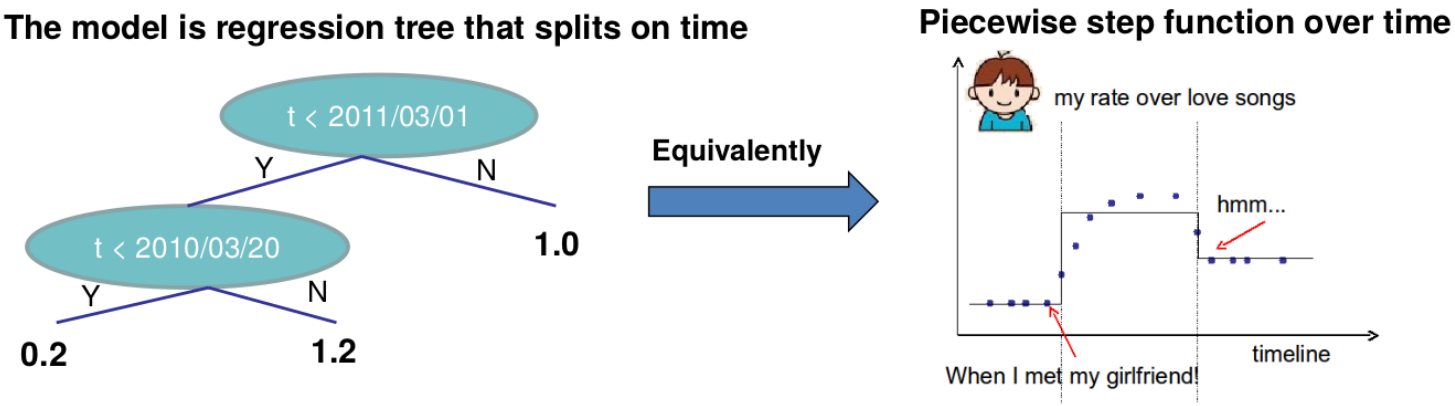
$$\hat{y} = \sum_{i}^{K} f_i(x_i), \quad f_i \in \mathcal{F}$$

onde $\mathcal{F}$ é o conjunto de todas as funções geradas por árvores regressoras

  - Agora, em vez de aprender os coeficientes, estamos aprendendo as funções descritas por árvores.

Árvores Regressoras

- Como podemos aprender as funções?
- Define objetivos (loss + regularização), e otimize!!
- Exemplo:
  - Considere uma árvore regressora em uma simples entrada
  - Eu quero predizer o gosto de música romântica no momento t



Objetivo para árvore regressora

- Modelo: assumindo que nós tenhamos K árvores

$$\hat{y} = \sum_{i}^{K} f_i(x_i), \quad f_i \in \mathcal{F}$$

onde $\mathcal{F}$ é o conjunto de todas as funções geradas por árvores regressoras

- Objetivo

$$Obj = \sum_{i}^{N} l(y_i, \hat{y}_i) + \sum_{k}^{K} \Omega(f_k)$$

- Possíveis modos de definir Ω
  - Número de nodos na árvore, profundidade
  - Norma l2 no peso das folhas

Objetivos vs Heurísticas

- Quando falamos de árvore, são utilizadas as seguintes heurísticas:
  - Dividir a árvore por ganho de informação
  - Cortar a árvore
  - Profundidade máxima
  - Suavizar os valores folhas
- Podemos fazer a mesma analogia para árvores regressoras
  - Dividir a árvore por ganho de informação – training loss
  - Cortar a árvore – regularização definidos pelo número de nodos
  - Profundidade máxima – restrição no espaço de funções
  - Suavizar os valores folhas – regularização l2 nos pesos dos nodos folhas

Entrada: Conj. Treino $\{(x_i, y_i)\}_{i=1}^{n}$,, loss function $L(y, F(x))$, num. iterações M.
Algoritmo:

1. Inicializa o modelo com valor constante:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

2. para m = 1 até M:
   1. Calcula os resíduos:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

   2. Ajusta a árvore $h_m(x)$ para os resíduos, i.e. treina usando o conjunto $\{(x_i, r_{im})\}_{i=1}^{n}$.
   3. Calc. param. γ_m resolvendo o problema:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

   4. Atualiza o modelo:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. Retorna $F_M(x)$.

SVM para Regressão

- O SVM também se baseia em regressor linear da forma

$$f(x) = w^T x + \epsilon$$

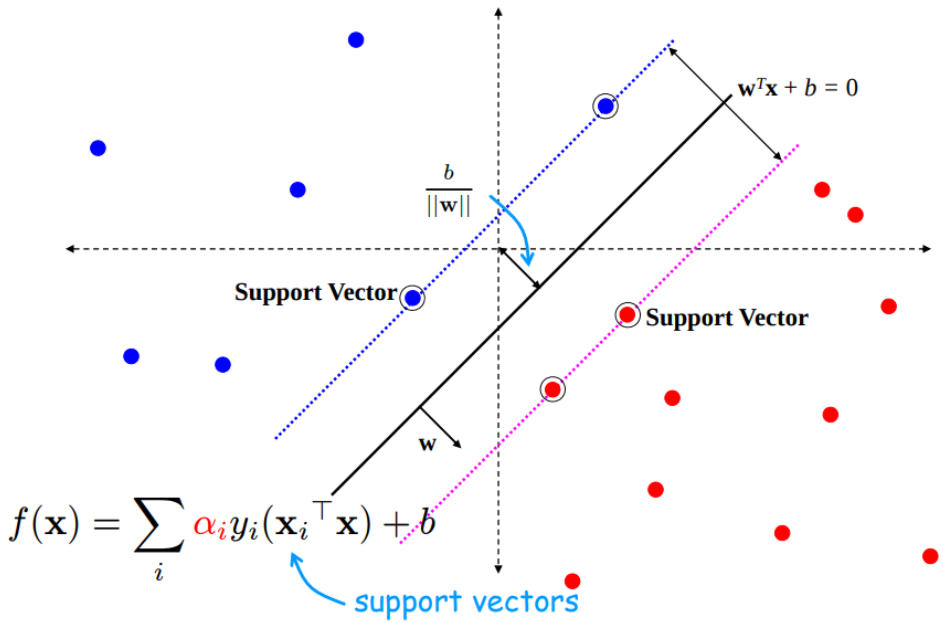
A formulação baseia-se na resolução do seguinte problema sobre w:

$$\min_{w \in \mathcal{R}} ||w||^2 + C \sum_{i}^{N} max(0, 1 - y_i f(x_i))$$

- A formulação é um problema de forma quadrática, pode-se escrever sua versão dual
- Assim, o SVM pode ser formulado para um problema linear da seguinte forma

$$f(x) = \sum_{i}^{N} \alpha_i y_i (x_i^T x) + \epsilon$$

que é otimizado para α


$\mathbf{w}^T\mathbf{x} + b = 0$
$\frac{b}{||\mathbf{w}||}$
Support Vector
Support Vector
$\mathbf{w}$
$f(\mathbf{x}) = \sum_i \alpha_i y_i (\mathbf{x}_i{}^\top \mathbf{x}) + b$
support vectors

Trabalho

Faça um jupyter notebook com a análise de regressão de um conjunto de dados. Neste trabalho você terá a liberdade de escolher o conjunto de dados.
Na análise dos dados, verá avaliado vários aspectos discutidos em aula. A seguir estão os itens avaliados no trabalho.

1. Definição e relevância do problema. Espera-se uma base de dados interessante, com vários exemplos e características. Bases triviais, bastante conhecidas, ou com análise já disponível, receberão baixo valor na avaliação
2. Faça a visualização dos dados. Entenda o conjunto de dados, suas distribuição, de forma visual. Faça plots adequados relacionando a variável resposta e as várias características do problema. Use os coeficientes de Pearson, ou de Spearman para verificar a correlação.
3. Descreva o pré-processamento realizado. Se foi retirado outliers, limpeza de dados, normalização.
4. Tente indicar manipulação de características, como remoção ou adição de novas, e como isso melhora na avaliação do modelo.
5. Aplique os modelos de Regressão (Regressão Linear, Ridge Regression, Lasso Regression, XGBoost, ElasticNet, KNN Regression, SVM Regression). Faça a otimização dos parâmetros dos regressores escolhidos utilizando o Grid Search. Descreva os melhores parâmetros encontrados.
6. (EXTRA) Faça um esquema de comitê (Ensemble) com o resultado dos vários modelos escolhidos. Veja se isso melhora os resultados.

Caso vários alunos apresentem a análise para um mesmo conjunto de dados, a maior nota será atribuído para aquele que obtiver o melhor score. Em caso de empate, ou seja, scores semelhantes para uma mesma base, tem-se a evidência de plágio. Neste caso os trabalhos serão analisados cuidadosamente para definir se houve ou não plágio.