Mineração de Dados

Aula 1: Introdução

Rafael Izbicki

Por que você está aqui?

- Porque você gosta do assunto
- Porque é um requerimento
- Porque Mineração de Dados é A matéria que vai te fazer ganhar $$$
- ...

Por que você está aqui?

- Porque você gosta do assunto
- Porque é um requerimento
- Porque Mineração de Dados é A matéria que vai te fazer ganhar $$$
- . . .

Por que você está aqui?

- Porque você gosta do assunto
- Porque é um requerimento
- Porque Mineração de Dados é A matéria que vai te fazer ganhar $$$
- ...

Por que você está aqui?

- Porque você gosta do assunto
- Porque é um requerimento
- Porque Mineração de Dados é A matéria que vai te fazer ganhar $$$
- ...

Por que você está aqui?

- Porque você gosta do assunto
- Porque é um requerimento
- Porque Mineração de Dados é A matéria que vai te fazer ganhar $$$
- . . .

Avaliação do Conhecimento Prévio

Estrutura do Curso

- Lembre-se de verificar o site da disciplina!
- 48 horas para responder os mensagens!

Estrutura do Curso

- Lembre-se de verificar o site da disciplina!
- 48 horas para responder os mensagens!

Estrutura do Curso: Material

- Slides em Aula

- Livro texto:

  Hastie T., Tibshirani R. And Friedman J. *The Elements of Statistical Learning*, Springer, 2009.

  James, G., Witten, D., Hastie, T. e Tibshirani, R. *An Introduction to Statistical Learning, with Applications in R*, Springer 2013.

  Disponível gratuitamente em
  `http://statweb.stanford.edu/~tibs/ElemStatLearn/`
  `http://www-bcf.usc.edu/~gareth/ISL/`

Estrutura do Curso: Material

- Slides em Aula

- Livro texto:

  Hastie T., Tibshirani R. And Friedman J. *The Elements of Statistical Learning*, Springer, 2009.

  James, G., Witten, D., Hastie, T. e Tibshirani, R. *An Introduction to Statistical Learning, with Applications in R*, Springer 2013.

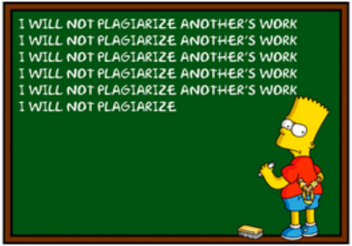
  Disponível gratuitamente em
  `http://statweb.stanford.edu/~tibs/ElemStatLearn/`
  `http://www-bcf.usc.edu/~gareth/ISL/`

Estrutura do Curso: Notas

⇒ Listas de Exercício (15% da nota): não deixe para a última hora! Não copie!

Para cada dia de atraso, pontos serão descontados

Duplas sorteadas dentro e fora da sala



Estrutura do Curso: Notas

⇒ Listas de Exercício (15% da nota): não deixe para a última hora! Não copie!

Para cada dia de atraso, pontos serão descontados

Duplas sorteadas dentro e fora da sala

Estrutura do Curso: Notas

$\Rightarrow$ Prova 1 (25% da nota)

$\Rightarrow$ Prova 2 (25% da nota)

$\Rightarrow$ Seminário (15% da nota)

$\Rightarrow$ Trabalho Prático (20% da nota)

Mais detalhes em breve

Estrutura do Curso: Notas

$\Rightarrow$ Prova 1 (25% da nota)

$\Rightarrow$ Prova 2 (25% da nota)

$\Rightarrow$ Seminário (15% da nota)

$\Rightarrow$ Trabalho Prático (20% da nota)

Mais detalhes em breve

Estrutura do Curso: Notas

$\Rightarrow$ Prova 1 (25% da nota)

$\Rightarrow$ Prova 2 (25% da nota)

$\Rightarrow$ Seminário (15% da nota)

$\Rightarrow$ Trabalho Prático (20% da nota)

Mais detalhes em breve

Estrutura do Curso: Notas

$\Rightarrow$ Prova 1 (25% da nota)

$\Rightarrow$ Prova 2 (25% da nota)

$\Rightarrow$ Seminário (15% da nota)

$\Rightarrow$ Trabalho Prático (20% da nota)

Mais detalhes em breve

Estrutura do Curso: Notas

$\Rightarrow$ Prova 1 (25% da nota)

$\Rightarrow$ Prova 2 (25% da nota)

$\Rightarrow$ Seminário (15% da nota)

$\Rightarrow$ Trabalho Prático (20% da nota)

Mais detalhes em breve

O que é o Aprendizado de Máquina?

Também conhecida como:

- Data Mining (Mineração de Dados)
- Big Data
- Inteligência Artificial

Aprendizado de Máquina é a ciência de descobrir estruturas e fazer predições em conjuntos de dados (grandes).

- Aprendizado supervisionado: Dadas medições $(X_1, Y_1), \ldots, (X_n, Y_n)$, aprender um modelo para prever Y_i baseado em X_i
- Aprendizado não supervisionado: Dadas medições $X_1, \ldots, X_n$, descobrir alguma estrutura baseado em similaridade

O que é o Aprendizado de Máquina?

Também conhecida como:

- Data Mining (Mineração de Dados)
- Big Data
- Inteligência Artificial

Aprendizado de Máquina é a ciência de descobrir estruturas e fazer predições em conjuntos de dados (grandes).

- Aprendizado supervisionado: Dadas medições $(X_1, Y_1), \ldots, (X_n, Y_n)$, aprender um modelo para prever Y_i baseado em X_i
- Aprendizado não supervisionado: Dadas medições $X_1, \ldots, X_n$, descobrir alguma estrutura baseado em similaridade

O que é o Aprendizado de Máquina?

Também conhecida como:

- Data Mining (Mineração de Dados)
- Big Data
- Inteligência Artificial

Aprendizado de Máquina é a ciência de descobrir estruturas e fazer predições em conjuntos de dados (grandes).

- Aprendizado supervisionado: Dadas medições $(X_1, Y_1), \ldots, (X_n, Y_n)$, aprender um modelo para prever Y_i baseado em X_i
- Aprendizado não supervisionado: Dadas medições $X_1, \ldots, X_n$, descobrir alguma estrutura baseado em similaridade

O que é o Aprendizado de Máquina?

Também conhecida como:

- Data Mining (Mineração de Dados)
- Big Data
- Inteligência Artificial

Aprendizado de Máquina é a ciência de descobrir estruturas e fazer predições em conjuntos de dados (grandes).

- Aprendizado supervisionado: Dadas medições $(X_1, Y_1), \ldots, (X_n, Y_n)$, aprender um modelo para prever Y_i baseado em X_i
- Aprendizado não supervisionado: Dadas medições $X_1, \ldots, X_n$, descobrir alguma estrutura baseado em similaridade

O que é o Aprendizado de Máquina?

Também conhecida como:

- Data Mining (Mineração de Dados)
- Big Data
- Inteligência Artificial

Aprendizado de Máquina é a ciência de descobrir estruturas e fazer predições em conjuntos de dados (grandes).

- Aprendizado supervisionado: Dadas medições $(X_1, Y_1), \ldots, (X_n, Y_n)$, aprender um modelo para prever Y_i baseado em X_i
- Aprendizado não supervisionado: Dadas medições $X_1, \ldots, X_n$, descobrir alguma estrutura baseado em similaridade

O que é o Aprendizado de Máquina?

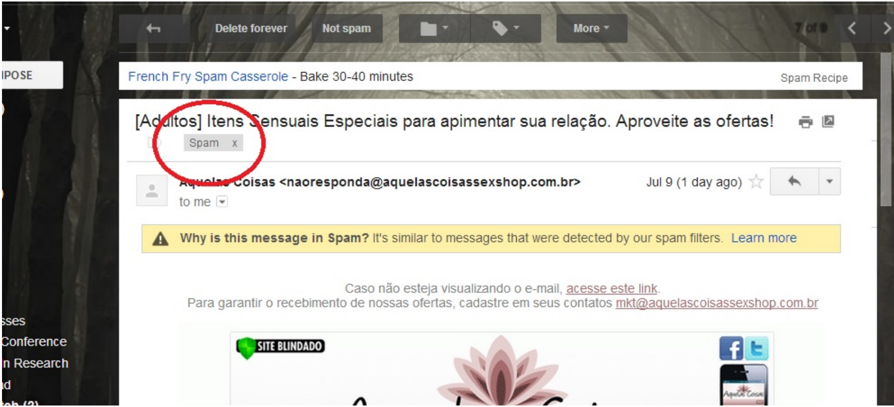
Também conhecida como:

- Data Mining (Mineração de Dados)
- Big Data
- Inteligência Artificial

Aprendizado de Máquina é a ciência de descobrir estruturas e fazer predições em conjuntos de dados (grandes).

- Aprendizado supervisionado: Dadas medições $(X_1, Y_1), \ldots, (X_n, Y_n)$, aprender um modelo para prever Y_i baseado em X_i
- Aprendizado não supervisionado: Dadas medições $X_1, \ldots, X_n$, descobrir alguma estrutura baseado em similaridade

Exemplo: Detecção de Spams

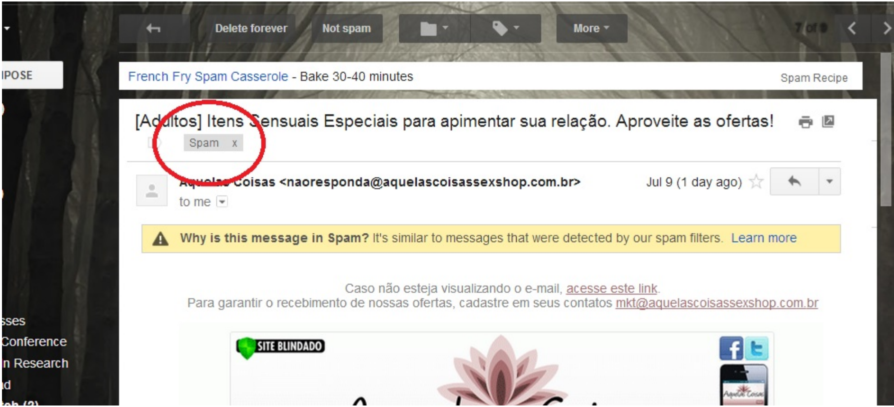


$X_i \longrightarrow$ email
$Y_i \longrightarrow$ spam/não spam

Objetivo: prever Y_i com base em X_i

Conjunto de dados grande!

Exemplo: Detecção de Spams

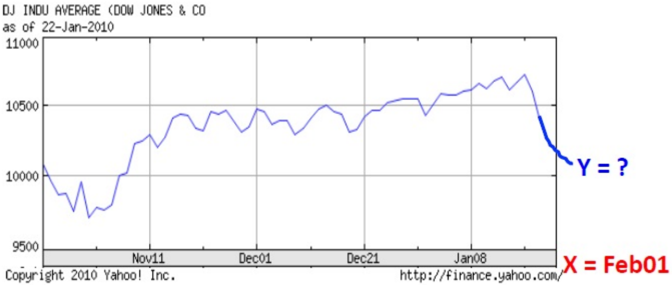


$X_i \longrightarrow$ email
$Y_i \longrightarrow$ spam/não spam

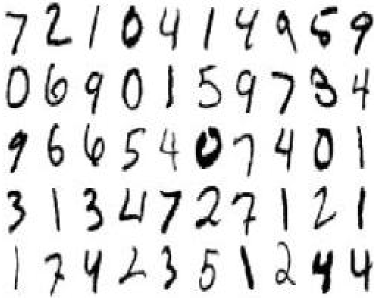
Objetivo: prever Y_i com base em X_i

Conjunto de dados grande!

Exemplo: Predição da Bolsa

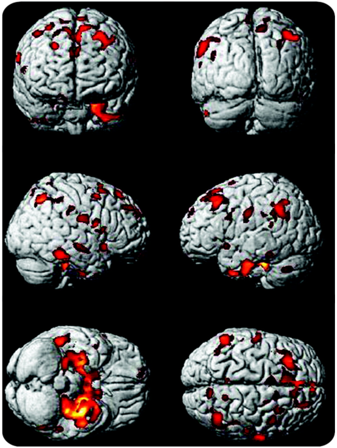


Exemplo: Reconhecimento de Dígitos



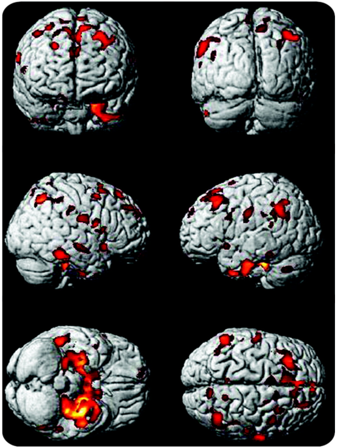
$X_i \longrightarrow$ imagem de um dígito
$Y_i \longrightarrow$ dígito correspondente

Exemplo: Predição de Alzheimer



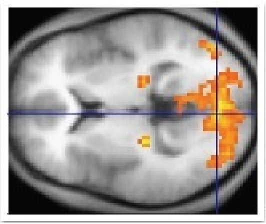
$X_i \longrightarrow$ imagem da ressonância magnética
$Y_i \longrightarrow$ Paciente com/sem Alzheimer

Exemplo: Predição de Alzheimer

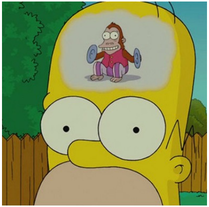


$X_i \longrightarrow$ imagem da ressonância magnética
$Y_i \longrightarrow$ Paciente com/sem Alzheimer

Exemplo: Leitura de Pensamentos

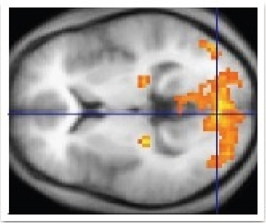


$\Downarrow\Downarrow\Downarrow$

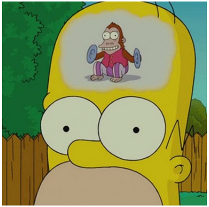


$X_i \longrightarrow$ imagem da ressonância magnética
$Y_i \longrightarrow$ "Pensamento"

Exemplo: Leitura de Pensamentos

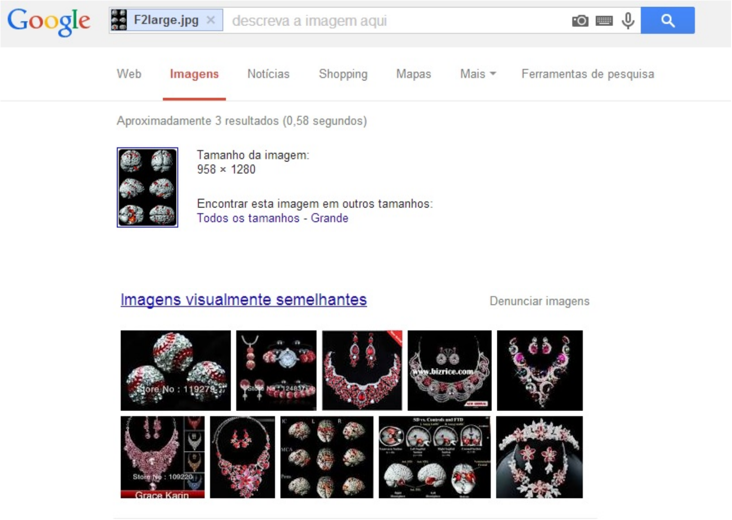


$\Downarrow\Downarrow\Downarrow$



$X_i \longrightarrow$ imagem da ressonância magnética
$Y_i \longrightarrow$ "Pensamento"

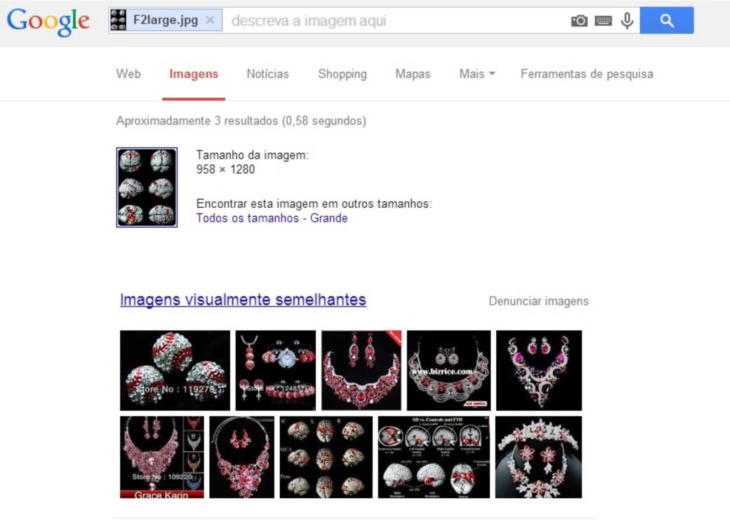
Exemplo: Busca por Imagens Semelhantes



$X_i \longrightarrow$ imagens na internet

Busca por estrutura

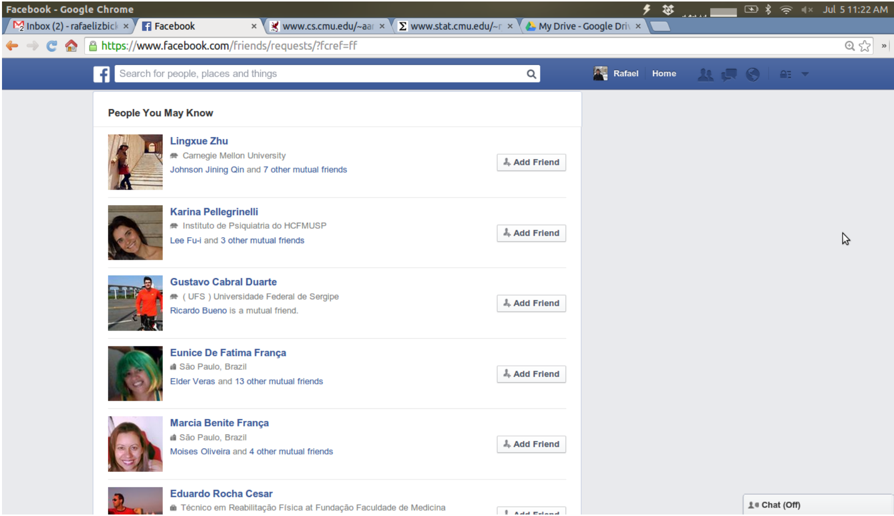
Exemplo: Busca por Imagens Semelhantes



$X_i \longrightarrow$ imagens na internet

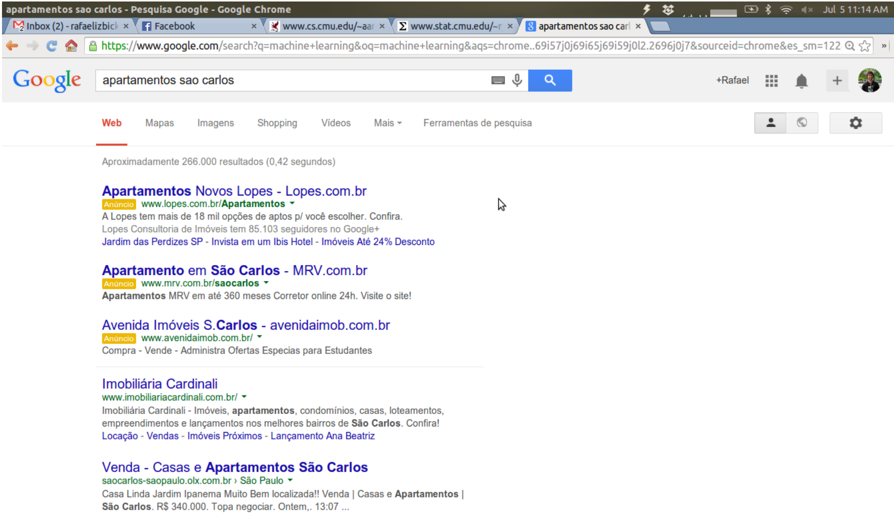
Busca por estrutura

Exemplo: Recomendação de Amizades



$X_i \longrightarrow$ existe um link entre dois usuários

Exemplo: Recomendação de Propagandas



Exemplo: Sistemas de Recomendação

Imagine que temos um banco de dados em que cada linha representa a ida de uma pessoa a um supermercado, e cada coluna representa se ela comprou ou não determinado produto.

Objetivo: descobrir regras do tipo

"Quem compra leite em geral também compra pão",

"Quem compra cerveja e refrigerante em geral também compra carne".

"Quem compra fralda em geral também compra cerveja".

Exemplo: Sistemas de Recomendação

Imagine que temos um banco de dados em que cada linha representa a ida de uma pessoa a um supermercado, e cada coluna representa se ela comprou ou não determinado produto.

Objetivo: descobrir regras do tipo

"Quem compra leite em geral também compra pão",

"Quem compra cerveja e refrigerante em geral também compra carne".

"Quem compra fralda em geral também compra cerveja".

Exemplo: Sistemas de Recomendação

Imagine que temos um banco de dados em que cada linha representa a ida de uma pessoa a um supermercado, e cada coluna representa se ela comprou ou não determinado produto.

Objetivo: descobrir regras do tipo

"Quem compra leite em geral também compra pão",

"Quem compra cerveja e refrigerante em geral também compra carne".

"Quem compra fralda em geral também compra cerveja".

Exemplo: Sistemas de Recomendação

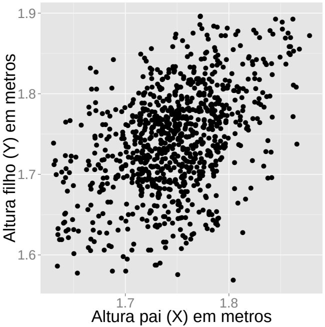
Aprendizado de Máquina é portanto uma ferramente extremamente útil hoje para fazer predições (*aprendizado supervisionado*) e aprender estruturas (*aprendizado não supervisionado*) em conjuntos de dados grandes

Vamos entrar mais fundo no problema de predição.

Aprendizado de Máquina é portanto uma ferramente extremamente útil hoje para fazer predições (*aprendizado supervisionado*) e aprender estruturas (*aprendizado não supervisionado*) em conjuntos de dados grandes

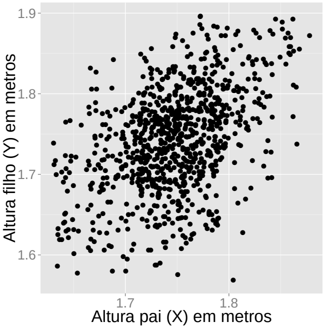
Vamos entrar mais fundo no problema de predição.

Vamos assumir que deseja-se prever a altura de um filho (Y) com base na altura de seu pai (X), e que coletamos uma amostra $(X_1, Y_1), \ldots, (X_n, Y_n)$ com observações sobre n indivíduos.



Com base nesses dados, podemos criar uma função de predição $g(x)$: dado que a altura do pai de um indivíduo é $X = x$, $g(x)$ representa nossa predição sobre a altura do filho Y.

Vamos assumir que deseja-se prever a altura de um filho (Y) com base na altura de seu pai (X), e que coletamos uma amostra $(X_1, Y_1), \ldots, (X_n, Y_n)$ com observações sobre n indivíduos.



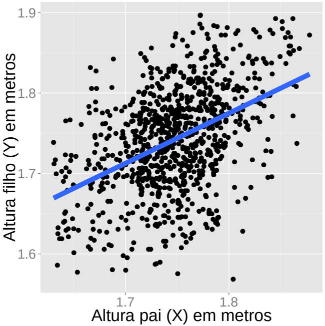
Com base nesses dados, podemos criar uma função de predição $g(x)$: dado que a altura do pai de um indivíduo é $X = x$, $g(x)$ representa nossa predição sobre a altura do filho Y.

Os principais objetivos de métodos de predição são (i) construir g de modo a se obter boas predições, e (ii) saber quantificar o quão boa uma (função de) predição é.

Uma maneira simples de se criar uma função de predição g

Os principais objetivos de métodos de predição são (i) construir g de modo a se obter boas predições, e (ii) saber quantificar o quão boa uma (função de) predição é.

Uma maneira simples de se criar uma função de predição g neste caso é através de uma regressão linear (vamos rever isto mais tarde).



Digamos que queremos predizer a altura de um filho cujo pai tem $x = 1,80$ m. Nossa predição neste caso é de $g(1,80) = 1,77$. Se observamos uma altura de 1,76 m, como quantificar quão errados estamos? (Exemplo no R)

Uma maneira de se fazer isso é usando-se o erro quadrático: neste caso, diriamos que o erro incorrido por g para prever $x = 1,80$ foi de

$$(g(x) - y)^2 = (1,76 - 1,77)^2 = 0.0001.$$

Quanto maior o valor de $(g(x) - y)^2$, pior é nossa predição.

Digamos que queremos predizer a altura de um filho cujo pai tem $x = 1,80$ m. Nossa predição neste caso é de $g(1,80) = 1,77$. Se observamos uma altura de 1,76 m, como quantificar quão errados estamos? (Exemplo no R)

Uma maneira de se fazer isso é usando-se o erro quadrático: neste caso, diriamos que o erro incorrido por g para prever $x = 1,80$ foi de

$$(g(x) - y)^2 = (1,76 - 1,77)^2 = 0.0001.$$

Quanto maior o valor de $(g(x) - y)^2$, pior é nossa predição.

Digamos que queremos predizer a altura de um filho cujo pai tem $x = 1,80$ m. Nossa predição neste caso é de $g(1,80) = 1,77$. Se observamos uma altura de 1,76 m, como quantificar quão errados estamos? (Exemplo no R)

Uma maneira de se fazer isso é usando-se o erro quadrático: neste caso, diriamos que o erro incorrido por g para prever $x = 1,80$ foi de

$$(g(x) - y)^2 = (1,76 - 1,77)^2 = 0.0001.$$

Quanto maior o valor de $(g(x) - y)^2$, pior é nossa predição.

Note que desta maneira quantificamos quão boa g é apenas para um dado par (x, y). Uma maneira de se generalizar isso (i.e., atribuir um único número quantificando quão boa g é) é através da função de risco $R(g)$:

$$R(g) = \mathbb{E}\left[(Y - g(X))^2\right].$$

Note que desta maneira quantificamos quão boa g é apenas para um dado par (x, y). Uma maneira de se generalizar isso (i.e., atribuir um único número quantificando quão boa g é) é através da função de risco $R(g)$:

$$R(g) = \mathbb{E}\left[(Y - g(X))^2\right].$$

$R(g)$ que avalia o erro preditivo de g. Por quê? Digamos que observamos um novo conjunto $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$.

Pela lei dos grandes números sabemos que, se m é grande,

$$\frac{1}{m}\sum_{i=1}^{m}(Y_{n+i} - g(X_{n+i}))^2 \approx R(g) := \mathbb{E}\left[(Y - g(X))^2\right]$$

Em outras palavras, $R(g)$ é aproximadamente uma média do erro quadrático para cada uma das novas observações.

Assim, se conseguimos g tal que $R(g)$ seja pequeno, teremos boas predições em novas observações.

$R(g)$ que avalia o erro preditivo de g. Por quê? Digamos que observamos um novo conjunto $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$.

Pela lei dos grandes números sabemos que, se m é grande,

$$\frac{1}{m}\sum_{i=1}^{m}(Y_{n+i} - g(X_{n+i}))^2 \approx R(g) := \mathbb{E}\left[(Y - g(X))^2\right]$$

Em outras palavras, $R(g)$ é aproximadamente uma média do erro quadrático para cada uma das novas observações.

Assim, se conseguimos g tal que $R(g)$ seja pequeno, teremos boas predições em novas observações.

$R(g)$ que avalia o erro preditivo de g. Por quê? Digamos que observamos um novo conjunto $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$.

Pela lei dos grandes números sabemos que, se m é grande,

$$\frac{1}{m}\sum_{i=1}^{m}(Y_{n+i} - g(X_{n+i}))^2 \approx R(g) := \mathbb{E}\left[(Y - g(X))^2\right]$$

Em outras palavras, $R(g)$ é aproximadamente uma média do erro quadrático para cada uma das novas observações.

Assim, se conseguimos g tal que $R(g)$ seja pequeno, teremos boas predições em novas observações.

$R(g)$ que avalia o erro preditivo de g. Por quê? Digamos que observamos um novo conjunto $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$.

Pela lei dos grandes números sabemos que, se m é grande,

$$\frac{1}{m}\sum_{i=1}^{m}(Y_{n+i} - g(X_{n+i}))^2 \approx R(g) := \mathbb{E}\left[(Y - g(X))^2\right]$$

Em outras palavras, $R(g)$ é aproximadamente uma média do erro quadrático para cada uma das novas observações.

Assim, se conseguimos g tal que $R(g)$ seja pequeno, teremos boas predições em novas observações.

$R(g)$ que avalia o erro preditivo de g. Por quê? Digamos que observamos um novo conjunto $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$.

Pela lei dos grandes números sabemos que, se m é grande,

$$\frac{1}{m}\sum_{i=1}^{m}(Y_{n+i} - g(X_{n+i}))^2 \approx R(g) := \mathbb{E}\left[(Y - g(X))^2\right]$$

Em outras palavras, $R(g)$ é aproximadamente uma média do erro quadrático para cada uma das novas observações.

Assim, se conseguimos g tal que $R(g)$ seja pequeno, teremos boas predições em novas observações.

Resumindo até agora

Em um problema de predição:

- Observamos um conjunto de treinamento $(X_1, Y_1), \ldots, (X_n, Y_n)$. X são chamados de preditores, ou variáveis explicativas, ou variáveis independentes, ou *features*. Y é chamado de resposta, ou variável dependente, ou *labels*
- Desejamos criar uma função de predição $g(x)$ para prever novas observações $X_{n+1}, \ldots, X_{n+m}$ bem
- Prever novas observações bem significa criar g tal que o risco quadrático $R(g) = \mathbb{E}\left[(Y - g(X))^2\right]$ seja baixo (dependendo da situação, outras funções de risco podem ser mais adequadas; veremos mais sobre isso depois)

Vamos ver ao longo do curso como, em diferentes contextos, criar g tal que $R(g)$ seja baixo, e como comparar diferentes g's.

Resumindo até agora

Em um problema de predição:

- Observamos um conjunto de treinamento $(X_1, Y_1), \ldots, (X_n, Y_n)$. X são chamados de preditores, ou variáveis explicativas, ou variáveis independentes, ou *features*. Y é chamado de resposta, ou variável dependente, ou *labels*
- Desejamos criar uma função de predição $g(x)$ para prever novas observações $X_{n+1}, \ldots, X_{n+m}$ bem
- Prever novas observações bem significa criar g tal que o risco quadrático $R(g) = \mathbb{E}\left[(Y - g(X))^2\right]$ seja baixo (dependendo da situação, outras funções de risco podem ser mais adequadas; veremos mais sobre isso depois)

Vamos ver ao longo do curso como, em diferentes contextos, criar g tal que $R(g)$ seja baixo, e como comparar diferentes g's.

Resumindo até agora

Em um problema de predição:

- Observamos um conjunto de treinamento $(X_1, Y_1), \ldots, (X_n, Y_n)$. X são chamados de preditores, ou variáveis explicativas, ou variáveis independentes, ou *features*. Y é chamado de resposta, ou variável dependente, ou *labels*
- Desejamos criar uma função de predição $g(x)$ para prever novas observações $X_{n+1}, \ldots, X_{n+m}$ bem
- Prever novas observações bem significa criar g tal que o risco quadrático $R(g) = \mathbb{E}\left[(Y - g(X))^2\right]$ seja baixo (dependendo da situação, outras funções de risco podem ser mais adequadas; veremos mais sobre isso depois)

Vamos ver ao longo do curso como, em diferentes contextos, criar g tal que $R(g)$ seja baixo, e como comparar diferentes g's.

Resumindo até agora

Em um problema de predição:

- Observamos um conjunto de treinamento $(X_1, Y_1), \ldots, (X_n, Y_n)$. X são chamados de preditores, ou variáveis explicativas, ou variáveis independentes, ou *features*. Y é chamado de resposta, ou variável dependente, ou *labels*
- Desejamos criar uma função de predição $g(x)$ para prever novas observações $X_{n+1}, \ldots, X_{n+m}$ bem
- Prever novas observações bem significa criar g tal que o risco quadrático $R(g) = \mathbb{E}\left[(Y - g(X))^2\right]$ seja baixo (dependendo da situação, outras funções de risco podem ser mais adequadas; veremos mais sobre isso depois)

Vamos ver ao longo do curso como, em diferentes contextos, criar g tal que $R(g)$ seja baixo, e como comparar diferentes g's.

Qual a melhor função $g(x)$?

O problema de encontrar g que minimize $R(g) = \mathbb{E}\left[(Y - g(X))^2\right]$ possui solução analítica!!

Seja $r(x) = \mathbb{E}[Y|X = x]$ a função de regressão (note que não estamos assumindo que ela é linear).

Então $R(r) \leq R(g)$ para toda função $g(x)$. Este é um exercício da lista.

O problema então está resolvido? Não!! $r(x) = \mathbb{E}[Y|X = x]$ depende de quantidades desconhecidas!

Contudo, este resultado indica que uma estimativa boa da função de regressão r nos fornece um bom preditor para y com base em x.

Qual a melhor função $g(x)$?

O problema de encontrar g que minimize $R(g) = \mathbb{E}\left[(Y - g(X))^2\right]$ possui solução analítica!!

Seja $r(x) = \mathbb{E}[Y|X = x]$ a função de regressão (note que não estamos assumindo que ela é linear).

Então $R(r) \leq R(g)$ para toda função $g(x)$. Este é um exercício da lista.

O problema então está resolvido? Não!! $r(x) = \mathbb{E}[Y|X = x]$ depende de quantidades desconhecidas!

Contudo, este resultado indica que uma estimativa boa da função de regressão r nos fornece um bom preditor para y com base em x.

Qual a melhor função $g(x)$?

O problema de encontrar g que minimize $R(g) = \mathbb{E}\left[(Y - g(X))^2\right]$ possui solução analítica!!

Seja $r(x) = \mathbb{E}[Y|X = x]$ a função de regressão (note que não estamos assumindo que ela é linear).

Então $R(r) \leq R(g)$ para toda função $g(x)$. Este é um exercício da lista.

O problema então está resolvido? Não!! $r(x) = \mathbb{E}[Y|X = x]$ depende de quantidades desconhecidas!

Contudo, este resultado indica que uma estimativa boa da função de regressão r nos fornece um bom preditor para y com base em x.

Qual a melhor função $g(x)$?

O problema de encontrar g que minimize $R(g) = \mathbb{E}\left[(Y - g(X))^2\right]$ possui solução analítica!!

Seja $r(x) = \mathbb{E}[Y|X = x]$ a função de regressão (note que não estamos assumindo que ela é linear).

Então $R(r) \leq R(g)$ para toda função $g(x)$. Este é um exercício da lista.

O problema então está resolvido? Não!! $r(x) = \mathbb{E}[Y|X = x]$ depende de quantidades desconhecidas!

Contudo, este resultado indica que uma estimativa boa da função de regressão r nos fornece um bom preditor para y com base em x.

Qual a melhor função $g(x)$?

O problema de encontrar g que minimize $R(g) = \mathbb{E}\left[(Y - g(X))^2\right]$ possui solução analítica!!

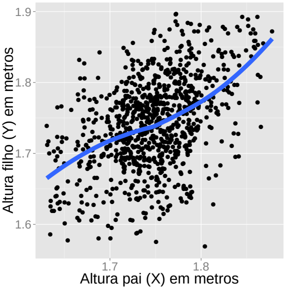
Seja $r(x) = \mathbb{E}[Y|X = x]$ a função de regressão (note que não estamos assumindo que ela é linear).

Então $R(r) \leq R(g)$ para toda função $g(x)$. Este é um exercício da lista.

O problema então está resolvido? Não!! $r(x) = \mathbb{E}[Y|X = x]$ depende de quantidades desconhecidas!

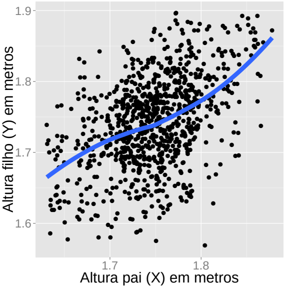
Contudo, este resultado indica que uma estimativa boa da função de regressão r nos fornece um bom preditor para y com base em x.

Uma forma de se criar/estimar g é através de uma regressão linear. Ao longo das próximas aulas veremos algumas metodologias mais sofisticadas para isso; inclusive métodos que podem ser aplicados quando x consiste em imagens, textos etc. Veja por exemplo um método mais sofisticado (não linear) para os dados das idades:



Antes disso, precisamos entender mais a fundo os elementos de um problema de predição.

Uma forma de se criar/estimar g é através de uma regressão linear. Ao longo das próximas aulas veremos algumas metodologias mais sofisticadas para isso; inclusive métodos que podem ser aplicados quando x consiste em imagens, textos etc. Veja por exemplo um método mais sofisticado (não linear) para os dados das idades:



Antes disso, precisamos entender mais a fundo os elementos de um problema de predição.

Regressão linear sob um ponto de vista preditivo

Regressão linear é (ou ao menos deveria ser) um velho conhecido seu.

É um estimator de $r(x) = \mathbb{E}[Y|X = x]$ quando ele assume uma forma linear, i.e.,

$$r(x) = \beta_0 + \beta_1 x,$$

ou, no caso em que $x = (x_1, \ldots, x_p)$ é um vetor,

$$r(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Em notação matricial,

$$r(x) = \beta^t x,$$

com $\beta = (\beta_0, \ldots, \beta_p)$ e $x = (x_1, \ldots, x_n)$

Regressão linear sob um ponto de vista preditivo

Regressão linear é (ou ao menos deveria ser) um velho conhecido seu.

É um estimator de $r(x) = \mathbb{E}[Y|X = x]$ quando ele assume uma forma linear, i.e.,

$$r(x) = \beta_0 + \beta_1 x,$$

ou, no caso em que $x = (x_1, \ldots, x_p)$ é um vetor,

$$r(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Em notação matricial,

$$r(x) = \beta^t x,$$

com $\beta = (\beta_0, \ldots, \beta_p)$ e $x = (x_1, \ldots, x_n)$

Regressão linear sob um ponto de vista preditivo

Regressão linear é (ou ao menos deveria ser) um velho conhecido seu.

É um estimator de $r(x) = \mathbb{E}[Y|X = x]$ quando ele assume uma forma linear, i.e.,

$$r(x) = \beta_0 + \beta_1 x,$$

ou, no caso em que $x = (x_1, \ldots, x_p)$ é um vetor,

$$r(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Em notação matricial,

$$r(x) = \beta^t x,$$

com $\beta = (\beta_0, \ldots, \beta_p)$ e $x = (x_1, \ldots, x_n)$

Regressão linear sob um ponto de vista preditivo

Regressão linear é (ou ao menos deveria ser) um velho conhecido seu.

É um estimator de $r(x) = \mathbb{E}[Y|X = x]$ quando ele assume uma forma linear, i.e.,

$$r(x) = \beta_0 + \beta_1 x,$$

ou, no caso em que $x = (x_1, \ldots, x_p)$ é um vetor,

$$r(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Em notação matricial,

$$r(x) = \beta^t x,$$

com $\beta = (\beta_0, \ldots, \beta_p)$ e $x = (x_1, \ldots, x_n)$

Regressão linear sob um ponto de vista preditivo

Relembrando: o estimador para os parâmetros β usual é o estimador de mínimos quadrados que em sua forma matricial é dado por

$$\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p) = (X^t X)^{-1} X^t Y.$$

Este é o estimador que minimiza

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p})^2$$

Assim, uma estimativa para $r(x) = \mathbb{E}[Y|X = x]$ é dada por

$$\widehat{r}(x) = \widehat{\beta}^t x.$$

Regressão linear sob um ponto de vista preditivo

Relembrando: o estimador para os parâmetros β usual é o estimador de mínimos quadrados que em sua forma matricial é dado por

$$\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p) = (X^t X)^{-1} X^t Y.$$

Este é o estimador que minimiza

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p})^2$$

Assim, uma estimativa para $r(x) = \mathbb{E}[Y|X = x]$ é dada por

$$\widehat{r}(x) = \widehat{\beta}^t x.$$

Regressão linear sob um ponto de vista preditivo

Relembrando: o estimador para os parâmetros β usual é o estimador de mínimos quadrados que em sua forma matricial é dado por

$$\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p) = (X^t X)^{-1} X^t Y.$$

Este é o estimador que minimiza

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p})^2$$

Assim, uma estimativa para $r(x) = \mathbb{E}[Y|X = x]$ é dada por

$$\widehat{r}(x) = \widehat{\beta}^t x.$$

Regressão linear sob um ponto de vista preditivo

No R, podemos usar a função `lm`:

ajuste=lm("y $\sim$ x",data=dados);

também é possível usar a forma matricial

Predição versus Inferência:

Em inferência em geral assume-se que o modelo é correto. Isso occore pois o principal objetivo está na interpretação dos parâmetros β. Ex: quais parâmetros são significantes? Qual o efeito do aumento da dose do remédio no medicamento? etc.

Já em predição, nosso objetivo maior é simplesmente criar $g(x)$ que tenha bom poder preditivo. Não estamos assumindo que a verdadeira regressão é linear!! Isto não quer dizer que não possamos interpretar nossos resultados, mas este em geral não é o foco das análises.

Predição versus Inferência:

Em inferência em geral assume-se que o modelo é correto. Isso occore pois o principal objetivo está na interpretação dos parâmetros β. Ex: quais parâmetros são significantes? Qual o efeito do aumento da dose do remédio no medicamento? etc.

Já em predição, nosso objetivo maior é simplesmente criar $g(x)$ que tenha bom poder preditivo. Não estamos assumindo que a verdadeira regressão é linear!! Isto não quer dizer que não possamos interpretar nossos resultados, mas este em geral não é o foco das análises.

As duas culturas

L. Breiman: Statistical modeling: The two cultures. Statistical Science, 16(3):199–231, 2001.

Duas culturas no uso de modelos estatísticos:

- Data Modeling Culture:

- Algorithmic Modeling Culture:

As duas culturas

L. Breiman: Statistical modeling: The two cultures. Statistical Science, 16(3):199–231, 2001.

Duas culturas no uso de modelos estatísticos:

- Data Modeling Culture:

- Algorithmic Modeling Culture:

As duas culturas

L. Breiman: Statistical modeling: The two cultures. Statistical Science, 16(3):199–231, 2001.

Duas culturas no uso de modelos estatísticos:

- Data Modeling Culture: Domina a comunidade estatística. Nela se assume que o modelo para $r(\vec{x})$ é correto. Testar suposições é fundamental. Foco em inferência.
- Algorithmic Modeling Culture:

As duas culturas

L. Breiman: Statistical modeling: The two cultures. Statistical Science, 16(3):199–231, 2001.

Duas culturas no uso de modelos estatísticos:

- Data Modeling Culture: Domina a comunidade estatística. Nela se assume que o modelo para $r(\vec{x})$ é correto. Testar suposições é fundamental. Foco em inferência.

- Algorithmic Modeling Culture: Domina a comunidade de machine learning. Nela **não** se assume que o modelo utilizado para $r(\vec{x})$ é correto; o modelo é utilizado apenas para criar bons algoritmos preditivos.

"Oddly, we are in a period where there has never been such a wealth of new statistical problems and sources of data. The danger is that if we define the boundaries of our field in terms of familar tools and familar problems, we will fail to grasp the new opportunities". (Breiman, 2001)

Resumo da Aula: Mineração de Dados é uma ferramente extremamente útil hoje para fazer predições (*aprendizado supervisionado*) e aprender estruturas (*aprendizado não supervisionado*) em conjuntos de dados grandes

Próxima Aula: Vamos entrar mais fundo no problema de predição

Resumo da Aula: Mineração de Dados é uma ferramenta extremamente útil hoje para fazer predições (*aprendizado supervisionado*) e aprender estruturas (*aprendizado não supervisionado*) em conjuntos de dados grandes

Próxima Aula: Vamos entrar mais fundo no problema de predição