

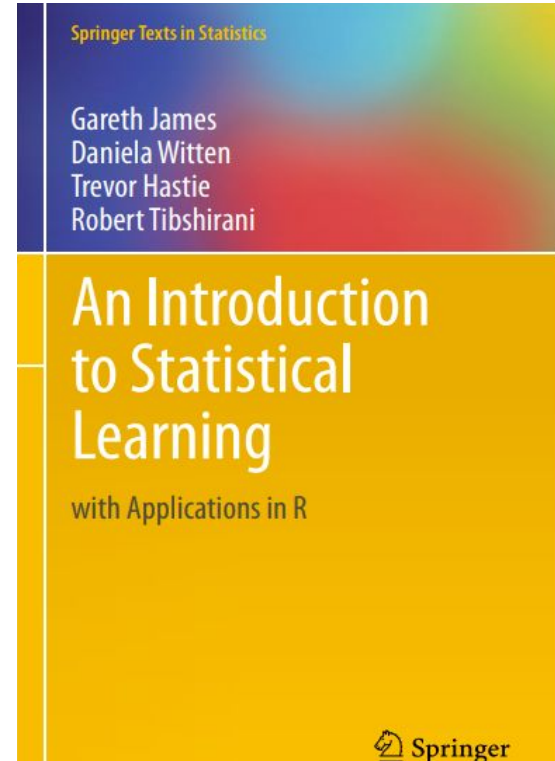
Cap 6: Linear Model Selection and Regularization

Walkiria Resende



Agenda

- Quem sou eu
- Revisão
- Terminologia
- Seleção de Variáveis
- Regularização
- Redução de Dimensionalidade
- Exercícios





Mineira de
São Tiago.
Nascida em
um sítio

Escola Rural /
Escola na
cidade



Graduada em
Ciência da
Computação -
2013

Mestre em
Bioinformática-
2017



Analista de
Sistemas e
DevOps
2013/2017

Analista de Validação
de modelos
Cientista de dados



*And mamãe do Dudu e esposa
do Henrique*



Aula 5: Regressão Linear - Priscila Portela

- <https://youtu.be/uxoitgv5FWU>
- https://www.youtube.com/watch?v=j1OYZJc9_RY



Terminologia

- Mínimos Quadrados: técnica para ajustar os parâmetros da regressão. Consistem em minimizar os **quadrados** dos resíduos.

- MSE (Mean Squared Error):
$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

- RMSE (Root Mean Squared Error) :

- MSPE (Mean Squared Percentual Error):
$$\text{MSPE} = \frac{100\%}{N} \sum_{i=1}^N \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2$$



Terminologia

- RSS (Residual Sum of Squares): $RSS = e_1^2 + e_2^2 + \dots + e_n^2$, $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$.
- TSS (Total Sum of Squares): $TSS = \sum (y_i - \bar{y})^2$



Terminologia

$$R^2 = 1 - \frac{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2} = 1 - \frac{MSE}{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2}$$

Ele nos dá uma idéia de quão bem podemos prever a variável resposta a partir da(s) variável(eis) preditor(a)s.

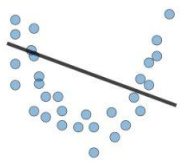


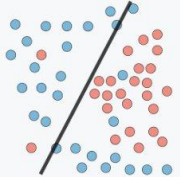
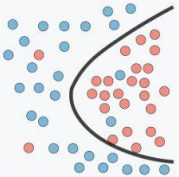
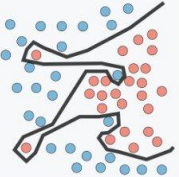
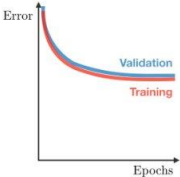
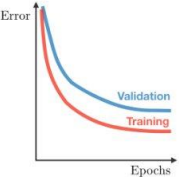
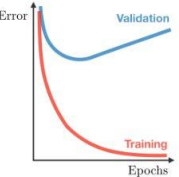
Ele representa a proporção da variabilidade na variável resposta explicada pela variável preditor(a) ou variável explanatória. Também conhecido como **coeficiente de determinação**.

1 se o modelo é perfeito, 0 caso contrário



Terminologia

- **Bias (Viés):** são os desvios entre aquilo que se observou no passado e aquilo que se prevê pelo modelo proposto – o modelo não tem toda a informação necessária para aprender
- **Variância:** modelo muito complexo que decorou os dados de treinamento

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none">• High training error• Training error close to test error• High bias	<ul style="list-style-type: none">• Training error slightly lower than test error	<ul style="list-style-type: none">• Very low training error• Training error much lower than test error• High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	<ul style="list-style-type: none">• Complexify model• Add more features• Train longer		<ul style="list-style-type: none">• Perform regularization• Get more data

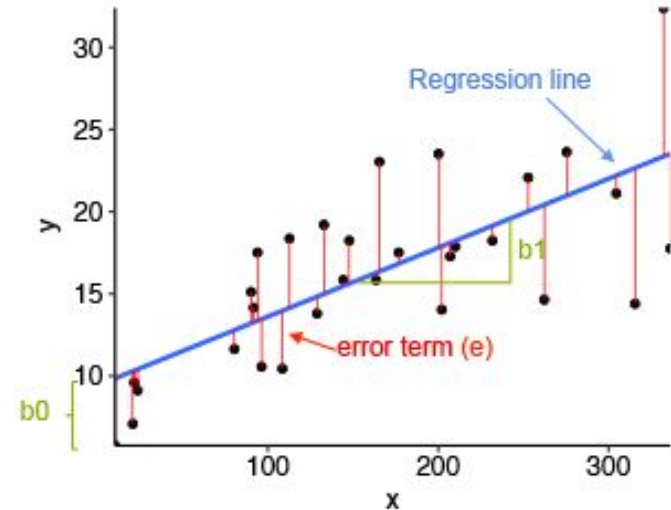
<https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks#classification-metrics>



Linear Model Selection and Regularization

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- B_0 = valor constante, intercepto da reta
- B_1 = inclinação da reta
- X = variáveis



Mínimos
Quadrados

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$



Mínimos Quadrados

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

y=Consumo	x=Renda
122	139
114	126
86	90
134	144
146	163
107	136
68	61
117	62
71	41
98	120
Média	
106,3	108,2



Mínimos Quadrados

$(y - y_{\text{medio}})$	$(x - x_{\text{medio}})$	$(y - y_{\text{medio}}) * (x - x_{\text{medio}})$	$(x - x_{\text{medio}})^2$
15,7	30,8	483,56	948,64
7,7	17,8	137,06	316,84
-20,3	-18,2	369,46	331,24
27,7	35,8	991,66	1281,64
39,7	54,8	2175,56	3003,04
0,7	27,8	19,46	772,84
-38,3	-47,2	1807,76	2227,84
10,7	-46,2	-494,34	2134,44
-35,3	-67,2	2372,16	4515,84
-8,3	11,8	-97,94	139,24



Mínimos Quadrados

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

soma((y-y_medio)*(x-x_medio))	soma((x-x_medio)2)
7764,4	15671,6

b1	0,4954439879
b0	52,6929605146
Consumo =	52,69 + 0,4954 x Renda + e

O aumento de R\$1,00 na renda causa um aumento de R\$0,49 no consumo



Linear Model Selection and Regularization

- Existem outras formas de aproximar os valores de β
- Mas porque faríamos de outra forma?

Acurácia

- Premissas:
 - Se a relação é realmente linear
 - Se $n \gg p$ (se o número de observações é maior que o número de variáveis)
- Mas se n não é muito maior que p , pode ocorrer overfitting
- E se $p > n$, o método de mínimos quadrados não pode ser usado,



Linear Model Selection and Regularization

Interpretabilidade

- Algumas variáveis associadas, não estão associadas a variável resposta e tornam o modelo mais complexo
- Existem técnicas que zeram os coeficientes dessas variáveis



Linear Model Selection and Regularization

Vamos discutir alguns métodos para realizar o ajuste da regressão

- **Seleção de um subconjunto de variáveis:** encontrar um conjunto de variáveis que nós acreditamos estar relacionado a variável resposta e então usar MQ para o fit
- **Shrinkage:** utiliza todas as variáveis para a predição, no entanto algumas serão reduzidas a zero ou muito próximos de zero. Esse método, também conhecido por regularização, reduz a variância. Também pode ser utilizado para seleção de variáveis
- **Redução de dimensionalidade:** projetar as variáveis em um espaço M , onde $M < p$ e depois usar MQ para o fit



Seleção de variáveis

All Features



Feature Selection



Final Features



Seleção de variáveis

- Incluir um grande número de covariáveis gera sérios problemas como interpretação complexa dos resultados
- As vezes é possível fazer uma **TRIAGEM INICIAL**, eliminando algumas variáveis antes da análise de regressão, se:
 - não são fundamentais para o estudo
 - são afetadas por grandes erros de medida
 - são redundantes com outra variável



Seleção de variáveis: Melhor subconjunto

Vazio

0,5



0,85



0,80



0,90

Vamos dar uma nota de 0 a 1 para esses alimentos!

Pensando em cada alimento, sem pensar em nenhum outro, sem comparação, o quanto você gosta de cada um?



Seleção de variáveis: Melhor subconjunto



0,75



0,6



0,98



0,95

Pensando nas combinações, sem pensar em outras combinações, o quanto você gosta de cada uma?



Seleção de variáveis: Melhor subconjunto

Comparando, o vencedor foi:



Seleção de variáveis: Melhor subconjunto

Pseudocódigo

1 Considere M_0 o modelo nulo, que não contém preditores. Este modelo prediz a média da amostra para cada observação

$$\binom{p}{k} = \frac{p!}{k!(p-k)!}$$

2 Para $k = 1, 2, \dots, p$:

(a) Ajuste todos os $\binom{p}{k}$ modelos que contêm exatamente k preditores

(b) Escolha o melhor entre os $\binom{p}{k}$ modelos, e chame-o de M_k . Aqui o melhor é definido usando o menor RSS ou equivalente, maior R^2

3 Selecione o melhor modelo entre M_0, M_1, \dots, M_p , usando erro da predição com validação cruzada, AIC, BIC ou R^2 ajustado



Seleção de variáveis: Melhor subconjunto

- No passo 2 do algoritmo, nós escolhemos o melhor modelo para cada subconjunto de variáveis.
- Varre todas as combinações possíveis de variáveis preditoras para encontrar o conjunto com melhor R^2 ajustado ou melhor critério C_p de Mallows, por exemplo
- Computacionalmente, pode ser bastante demandante, e pode se tornar inviável quando temos muitas variáveis candidatas
- Se tivermos M variáveis candidatas, há 2^M conjuntos possíveis; por exemplo, $M = 100$, há $1,268 \times 10^{30}$ regressões possíveis



Seleção de variáveis: Melhor subconjunto

Porque usar validação cruzada?



Seleção de variáveis: Melhor subconjunto

Porque usar validação cruzada?

- Em geral, a inclusão adicional de variáveis preditoras na regressão, como já mencionamos aumenta o coeficiente de determinação (R^2)
- A inclusão adicional de variáveis preditoras também reduz (mesmo que marginalmente) o erro de previsão (dentro da amostra) mesmo que as variáveis preditoras não façam sentido
- Portanto, quanto mais incluímos variáveis explicativas na regressão, o R^2 aumenta e a soma do quadrado dos erros diminui



Seleção de variáveis: Melhor subconjunto

Porque usar validação cruzada?

- O problema é que a soma dos quadrados dos erros SQE corresponde aos erros da regressão dentro da amostra (in-sample error)
- Nós gostaríamos de ter um modelo de regressão que possa ter boas previsões para dados fora da amostra
- Portanto, nós gostaríamos de ter um modelo que apresentasse baixo erro de previsão fora da amostra (out-of-sample error)
- Essa ideia de termos um bom modelo para previsão fora da amostra está intrinsecamente ligada aos procedimentos de validação cruzada (cross-validation) de um determinado modelo de regressão



Seleção de variáveis: Melhor subconjunto

Porque usar validação cruzada?

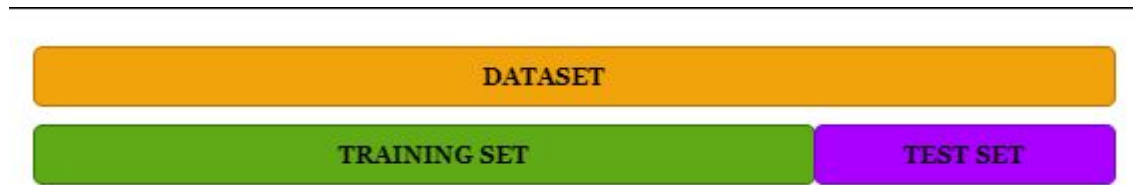
- Validação cruzada:
 - Dividimos a amostra em duas partes – podemos fazer uma divisão aleatória entre as observações que vão entrar em cada subamostra; a primeira amostra com n_1 observações e a segunda com n_2 observações
 - A primeira subamostra é usada para estimar os coeficientes da regressão ($\beta_0, \beta_1, \beta_2, \dots, \beta_k$)
 - Usamos os coeficientes estimados na primeira amostra para prever a variável resposta na segunda amostra
 - Calculamos agora o erro médio quadrático de previsão (mean square prediction error) com base apenas na segunda amostra



Seleção de variáveis: Melhor subconjunto

Porque usar validação cruzada?

- Validação cruzada:
 - Podemos então procurar o modelo de regressão, com as variáveis preditoras, que nos forneça o menor MSPE
 - O MSPE nos dá uma ideia do erro fora da amostra



Seleção de variáveis: Stepwise

- Limitações computacionais
- Overfitting

Forward, Backward, Híbridos



Seleção de variáveis: Forward Stepwise

Pseudocódigo

- 1 Considere M_0 o modelo nulo, que não contém preditores
- 2 Para $k = 1, 2, \dots, p-1$:
 - (a) Ajuste todos os $(p-k)$ modelos que aumenta os preditores em M_k com um preditor
 - (b) Escolha o melhor entre os $(p-k)$ modelos, e chame-o de M_{k+1} . Aqui o melhor é definido usando o menor RSS ou equivalente, maior R^2
- 3 Selecione o melhor modelo entre M_0, M_1, \dots, M_p , usando erro da predição com validação cruzada, AIC, BIC ou R^2 ajustado



Seleção de variáveis: Forward Stepwise

Essa é a quantidade de modelos gerados pela abordagem Forward:

$$\sum_{k=0}^{p-1} (p - k) = 1 + p(p + 1)/2$$

Se tivermos $p=20$, com a abordagem de Melhor Subconjunto, teremos 2^{20} , que gera 1,048,576, enquanto que na abordagem forward, teremos 211 modelos



Seleção de variáveis: Forward Stepwise

Vocês conseguem pensar em um problema com essa abordagem?



Seleção de variáveis: Forward Stepwise

Vocês conseguem pensar em um problema com essa abordagem?

Uma variável incluída, jamais será excluída! **Efeito nesting**



Seleção de variáveis: Forward Stepwise

Vocês conseguem pensar em um problema com essa abordagem?

Uma variável incluída, jamais será excluída! **Efeito nesting**

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income, student, limit	rating, income, student, limit



Seleção de variáveis: Forward Stepwise

Tarefinha:

Execute essa abordagem com o exemplo dos alimentos, feito com a abordagem de Melhor Modelo



Seleção de variáveis: Backward Stepwise

Pseudocódigo

- 1 Considere M_p o modelo cheio, que contém todos os preditores
- 2 Para $k=p, p-1, \dots, 1$:
 - (a) Considere todos os k modelos que contém todos os preditores, exceto 1, para todos os $k-1$
 - (b) Escolha o melhor entre os k modelos, e chame-o de M_{k-1} . Aqui o melhor é definido usando o menor RSS ou equivalente, maior R^2
- 3 Selecione o melhor modelo entre M_0, M_1, \dots, M_p , usando erro da predição com validação cruzada, AIC, BIC ou R^2 ajustado

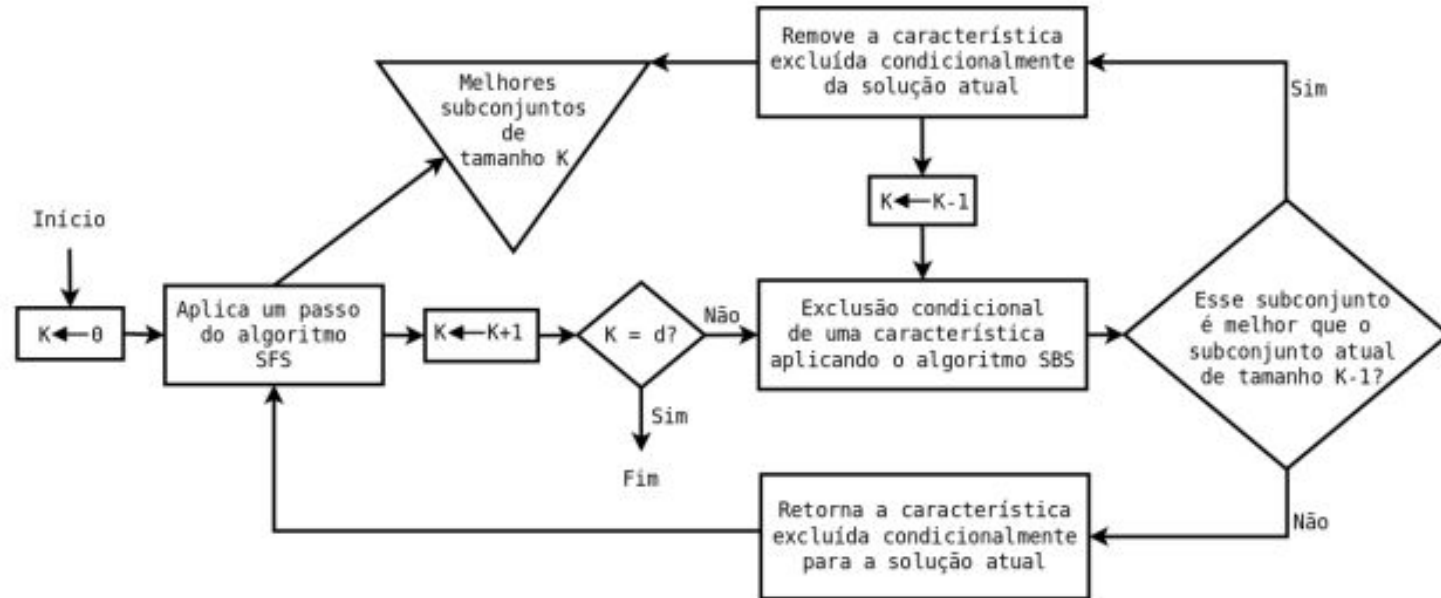


Seleção de variáveis: Backward Stepwise

- Nesse caso, **n** deve ser maior que **p**
 - para que o modelo full possa ser ajustado
- Em caso contrário, usar o forward
- Também sofre do efeito nesting



Seleção de variáveis: Híbrido



Escolhendo o melhor modelo



- As abordagens de melhor subconjunto ou stepwise, geram um conjunto de modelos, e precisamos compará-los, e então implementar o mais acurado
- Uma métrica muito utilizada em regressão é o R^2 e também o RSS
- Então podemos utilizar essas métricas para comparar os modelos?
- Infelizmente não podemos, porque eles não são ideais para comparar modelos com um número de preditores diferentes



Escolhendo o melhor modelo

1. Estimar o erro do conjunto de teste por meio de um ajuste do erro no conjunto de treinamento
2. Estimar o erro no conjunto de teste, utilizando validação cruzada



Escolhendo o melhor modelo

1. Estimar o erro do conjunto de teste por meio de um ajuste do erro no conjunto de treinamento:

$$C_p = \frac{1}{n} (RSS + 2d\hat{\sigma}^2),$$

d = número de preditores (ex.: M3 d=4)

σ^2 = erro do modelo full

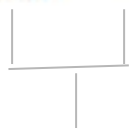
Se o submodelo com p parâmetros é melhor, então o valor do critério C_p deveria ser menor ou igual ao número de parâmetros p do modelo e aquele com menor valor do critério será um modelo relativamente preciso.*

Restrito a problemas com $n > p$



Escolhendo o melhor modelo

1. Estimar o erro do conjunto de teste por meio de um ajuste do erro no conjunto de treinamento:

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2} (\text{RSS} + 2d\hat{\sigma}^2),$$


-2logL

No caso da regressão linear, AIC e CP são equivalentes



Escolhendo o melhor modelo

1. Estimar o erro do conjunto de teste por meio de um ajuste do erro no conjunto de treinamento:

$$\text{BIC} = \frac{1}{n\hat{\sigma}^2} (\text{RSS} + \log(n)d\hat{\sigma}^2) .$$

Tende a escolher o modelo mais parcimonioso. Porque?



Escolhendo o melhor modelo

1. Estimar o erro do conjunto de teste por meio de um ajuste do erro no conjunto de treinamento:

$$\text{BIC} = \frac{1}{n\hat{\sigma}^2} (\text{RSS} + \log(n)d\hat{\sigma}^2).$$

Em AIC o termo é: $2(n)$

Tende a escolher o modelo mais parcimonioso. Porque?

Para qualquer $n > 7$, o valor de $\log n$ é maior que 2, logo a penalidade é maior



Escolhendo o melhor modelo

1. Estimar o erro do conjunto de teste por meio de um ajuste do erro no conjunto de treinamento:

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n - d - 1)}{\text{TSS}/(n - 1)}.$$

$$R^2 = 1 - \text{RSS}/\text{TSS}$$

$$\text{TSS} = \sum (y - y_m)^2$$

$$\text{RSS} = \sum (y - y_p)^2$$

O valor de R^2 ajustado não aumenta com a adição de variáveis



Escolhendo o melhor modelo

1. Estimar o erro do conjunto de teste por meio de um ajuste do erro no conjunto de treinamento:

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n - d - 1)}{\text{TSS}/(n - 1)}.$$

O valor de R^2 ajustado não aumenta com a adição de variáveis

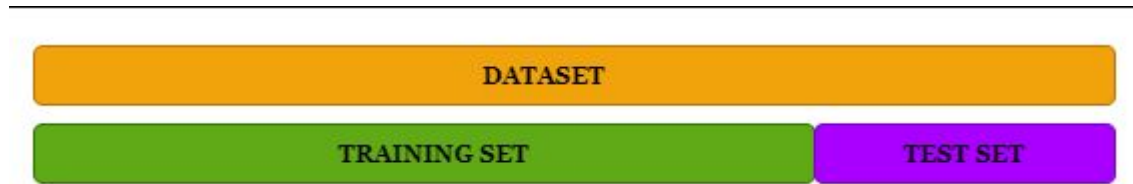
Vantagens: não precisamos estimar o σ^2 e pode ser aplicado em casos em que $p > n$



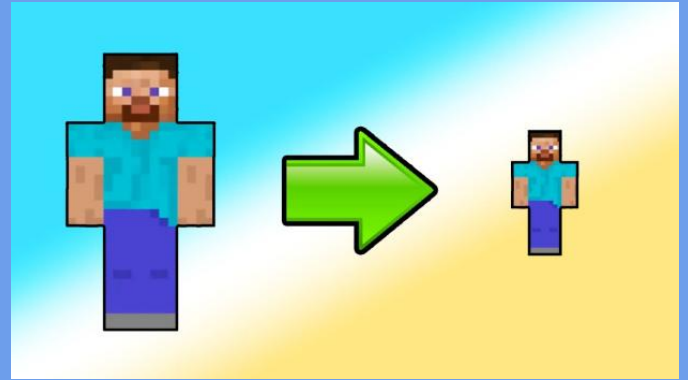
Escolhendo o melhor modelo

2. Estimar o erro no conjunto de teste, utilizando validação cruzada

Vantagem: não precisamos saber o valor de d , nem calcular o valor do erro para o modelo full

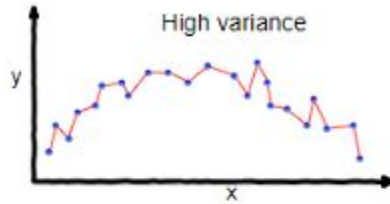


Shinkrage

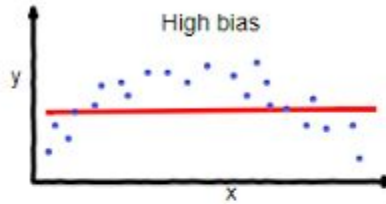


Shrinkage: Regularização

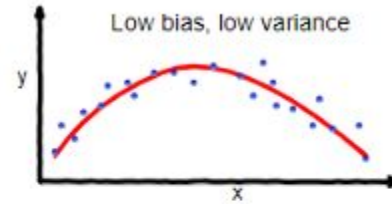
O que queremos?



overfitting



underfitting



Good balance



Shrinkage: Regularização

- Essa técnica desencoraja o ajuste excessivo dos dados, afim de diminuir a sua variância.
- **Ridge** e **Lasso** são formas de regularizarmos a nossa função através de penalidades. De forma simples, dentro de uma equação estatística dos dados, nós alteramos os fatores de forma a priorizar ou não certas parcelas da equação e, assim, evitamos 'overfitting' e melhoramos a qualidade de predição



Regularização: L2 Ridge

$$\text{RSS}_{\text{ridge}} = \sum_{i=1}^n [y_i - (\mathbf{w} \cdot \mathbf{x}_i + b)]^2 + \alpha \sum_{j=1}^p w_j^2$$

regularização ℓ_2

$\alpha == \lambda > 0$,

se $\lambda = 0$, então temos a regressão linear

quanto maior o valor de λ mais próximos de zero os coeficientes serão



Regularização: L2 Ridge

$$\text{RSS}_{\text{ridge}} = \sum_{i=1}^n [y_i - (\mathbf{w} \cdot \mathbf{x}_i + b)]^2 + \alpha \sum_{j=1}^p w_j^2$$

regularização ℓ_2

Vantagens:

- Computacional
- Funciona melhor em casos que a estimativa por MQ tem alta variância



Regularização: L2 Ridge

$$\text{RSS}_{\text{ridge}} = \sum_{i=1}^n [y_i - (\mathbf{w} \cdot \mathbf{x}_i + b)]^2 + \alpha \sum_{j=1}^p w_j^2$$

regularização ℓ_2

- Tarefa:
 - Em que casos MQ têm alta variância?



In general, in situations where the relationship between the response and the predictors is close to linear, the least squares estimates will have low bias but may have high variance. This means that a small change in the training data can cause a large change in the least squares coefficient estimates. In particular, when the number of variables p is almost as large as the number of observations n , as in the example in Figure 6.5, the least squares estimates will be extremely variable. And if $p > n$, then the least squares estimates do not even have a unique solution, whereas ridge regression can still perform well by trading off a small increase in bias for a large decrease in variance. Hence, ridge regression works best in situations where the least squares estimates have high variance.



Regularização: L1 Lasso

$$\text{RSS}_{\text{lasso}} = \sum_{i=1}^n [y_i - (\mathbf{w} \cdot \mathbf{x}_i + b)]^2, \quad \boxed{+ \alpha \sum_{j=1}^p |w_j|}$$

regularização ℓ_1

- Funciona como técnica de seleção de variáveis



Regularização: Comparação

- As elipses correspondem a função custo
- As áreas em azul correspondem a restrição

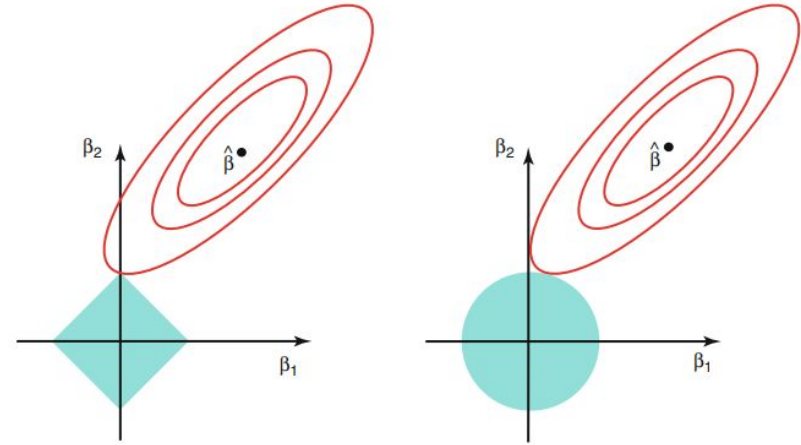


FIGURE 6.7. Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.



Regularização: Comparação

- Qual é melhor?
 - **DEPENDE**
- Normalmente, Lasso funciona bem quando temos um número menor de preditores
- O ideal é usar validação cruzada para encontrar a melhor técnica!



Regularização: Parâmetro de regularização

- A partir de uma grade de valores de λ , escolhemos aquele que resulta na menor taxa de erro de validação.
- Note que, quando $\lambda=0$, o termo de *penalty* não terá efeito na estimação por mínimos quadrados.
- A medida que $\lambda \rightarrow \infty$, a regularização aumenta.

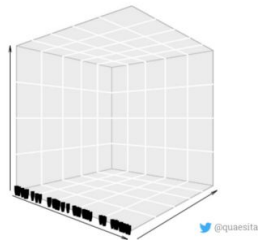


Redução de dimensionalidade

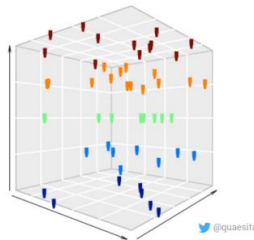
Mal da dimensionalidade

Mal da dimensionalidade:

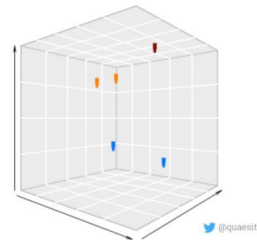
Quanto mais variáveis preditoras você tem mais dados você precisa!



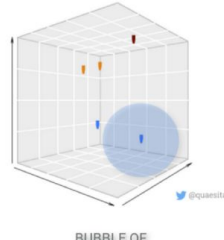
1D: LATITUDE



3D: LAT, LONG,
& FLOOR



11:00 AM



BUBBLE OF
LONELINESS

<https://towardsdatascience.com/the-curse-of-dimensionality-minus-the-curse-of-jargon-520da109fc87>



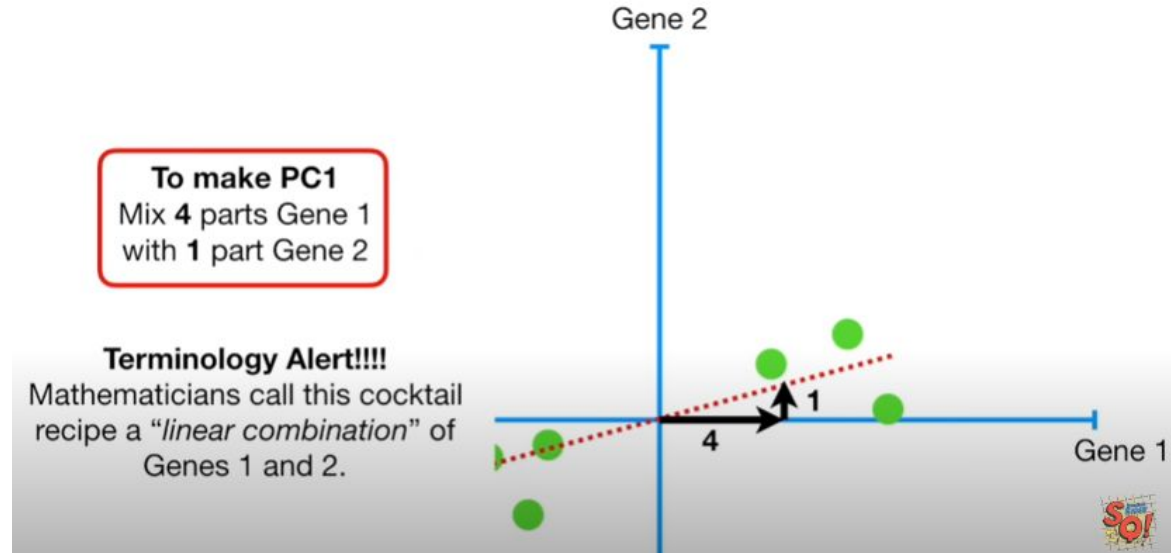
Redução de Dimensionalidade

- Os métodos anteriores se baseavam em selecionar um subconjunto de preditores e em zerar as variáveis. Mas as variáveis eram utilizadas no seu modo original
- Aqui nessa Seção, vamos transformar os preditores e então ajustar o modelo
- Os novos preditores serão uma combinação linear dos preditores originais



Redução de Dimensionalidade

- Os novos preditores serão uma combinação linear dos preditores originais:



Redução de Dimensionalidade

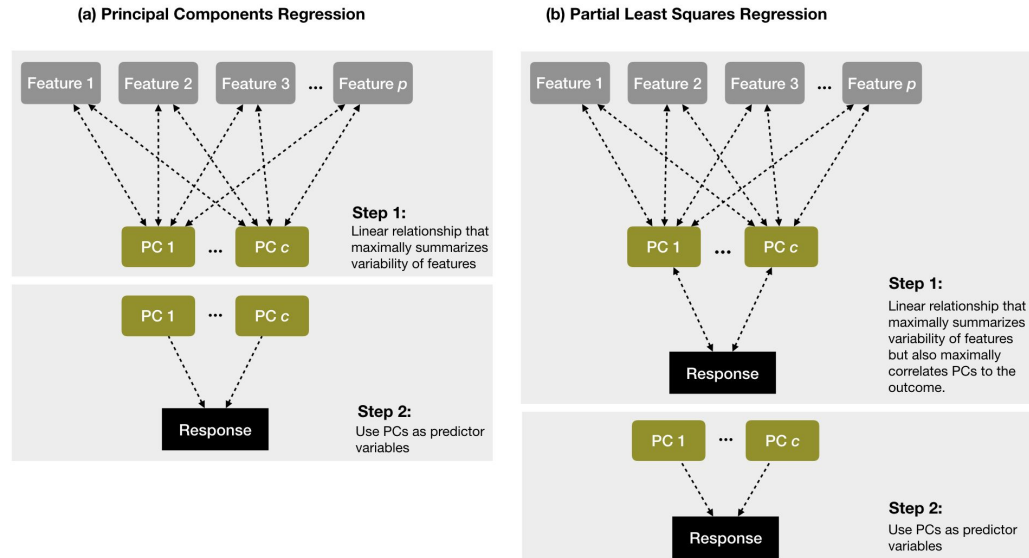
- Principal Components Regression
 - Obter os componentes principais.
 - Executar análise de regressão nos componentes principais.



Redução de Dimensionalidade

- Partial Least Squares

- usado quando as variáveis são altamente correlacionadas
- ou o número de variáveis é maior que o de observações



Redução de Dimensionalidade

Considerações finais:

“Ambas as técnicas criam novas variáveis independentes chamadas componentes que são combinações lineares das variáveis preditoras originais, mas a PCR cria componentes para explicar a variabilidade observada nas variáveis preditoras, sem considerar a variável resposta. Enquanto o PLS considera a variável dependente e, portanto, muitas vezes leva a modelos que são capazes de ajustar a variável dependente com menos componentes.” Ref:



Referências

<https://medium.com/the-owl/k-fold-cross-validation-in-keras-3ec4a3a00538>

[https://repositorio.enap.gov.br/bitstream/1/3452/12/Aula%207%20-%20Geraldo%20Goes%20e%20Alexandre%20Ywata%20-%200Selecao Variaveis v1.pdf](https://repositorio.enap.gov.br/bitstream/1/3452/12/Aula%207%20-%20Geraldo%20Goes%20e%20Alexandre%20Ywata%20-%200Selecao%20Variaveis%20v1.pdf)

<https://www.teses.usp.br/teses/disponiveis/95/95131/tde-27072011-105810/publico/tesefinalfabricio.pdf>

<http://library.utia.cas.cz/separaty/historie/somol-floating%20search%20methods%20in%20feature%20selection.pdf>

<https://medium.com/turing-talks/turing-talks-20-regress%C3%A3o-de-ridge-e-lasso-a0fc467b5629>

<https://www.kdnuggets.com/2019/10/feature-selection-beyond-feature-importance.html>

<http://www2.ic.uff.br/~aconci/PCA-ACP.pdf>

