Data Mining

LABDATA

Coordenação:

Prof. Dr. Adolpho Walter Pimazzi Canton

Profa. Dra. Alessandra de Ávila Montini

Disciplina: Machine Learning

Prof. Carlos Eduardo Martins Relvas

Currículo

- Bacharel em Estatística, Universidade de São Paulo.
- Mestre em Estatística, Universidade de São Paulo.
- Doutorando em Ciência da Computação, Universidade de São Paulo.

- Itaú, 2010-2015. Principais atividades:

  - Consultoria estatística para várias áreas do banco com foco principal em melhorias no processo de modelagem de risco de crédito.

  - De 2013 a 2015, participação do projeto Big Data do banco usando tecnologia Hadoop e diversas técnicas de machine learning. Desenvolvemos diversos algoritmos em MapReduce usando R e Hadoop streaming, criando uma plataforma de modelagem estatística no Hadoop.

- Nubank, dede 2015. Equipe de Data Science, responsável por toda a parte de modelagem da empresa, desde modelos de crédito e até mesmo identificar motivos de atendimento.

Agenda

- **Métricas**
- **Seleção de variáveis.**
- **Regularização**

Métricas

- Há diversas métricas para avaliar performance preditiva. Até o momento vimos o erro quadrático e erro absoluto para regressão e a acurácia para classificação.

- Cada métrica é adequada para uma situação específica e devemos saber quais os prós e contras de cada para sabermos qual métrica utilizar para o problema que estamos trabalhando.

- Veremos métricas específicas para regressão e específicas para classificação.

Regressão

Erro Absoluto Médio

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- Em que:

- y_i reflete o valor observado para a i-ésima variável.
- $\hat{y}_i$ representa o valor estimado para a i-ésima variável.

Raiz do Erro Quadrático Médio

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Métrica que mede a distância das previsões para os valores observados penalizando mais erros grandes.

R^2

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

- Em que:
- $\bar{y}$ representa a média da variável resposta

➢ Esta métrica compara a performance do modelo ao usarmos apenas a média como estimativa. Se for menor do que 0, significa que nosso modelo é pior do que usar a média. Se o modelo acertar todas observações, esta métrica será igual a 1.

Métricas Classificação

		Predicted condition		
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$
True condition	condition positive	**True positive**	**False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$
	condition negative	**False Positive** (Type I error)	**True negative**	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$
		Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	

Métricas Classificação

- F1 – score

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

- Varia entre 0 e 1.
- Quanto maior, melhor

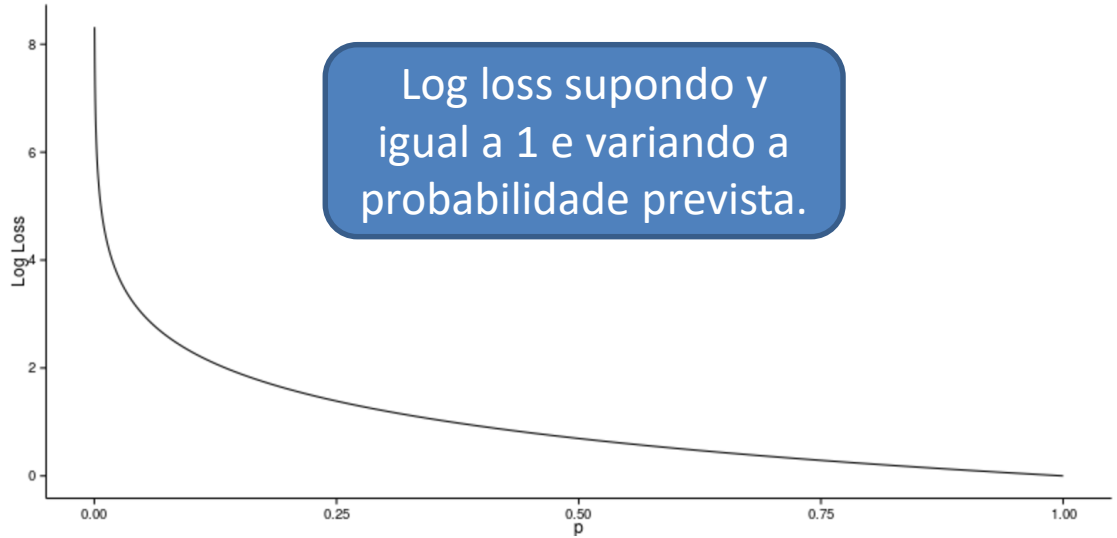
Acurácia

- Porcentagem de observações classificadas corretamente.

- Apresenta problemas em bases de dados com proporção de classes não balanceadas (por exemplo, 90% de bons e 10% de maus).

Log Loss

$$-\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1-y_i)\log(1-p_i)].$$



Log loss supondo y igual a 1 e variando a probabilidade prevista.

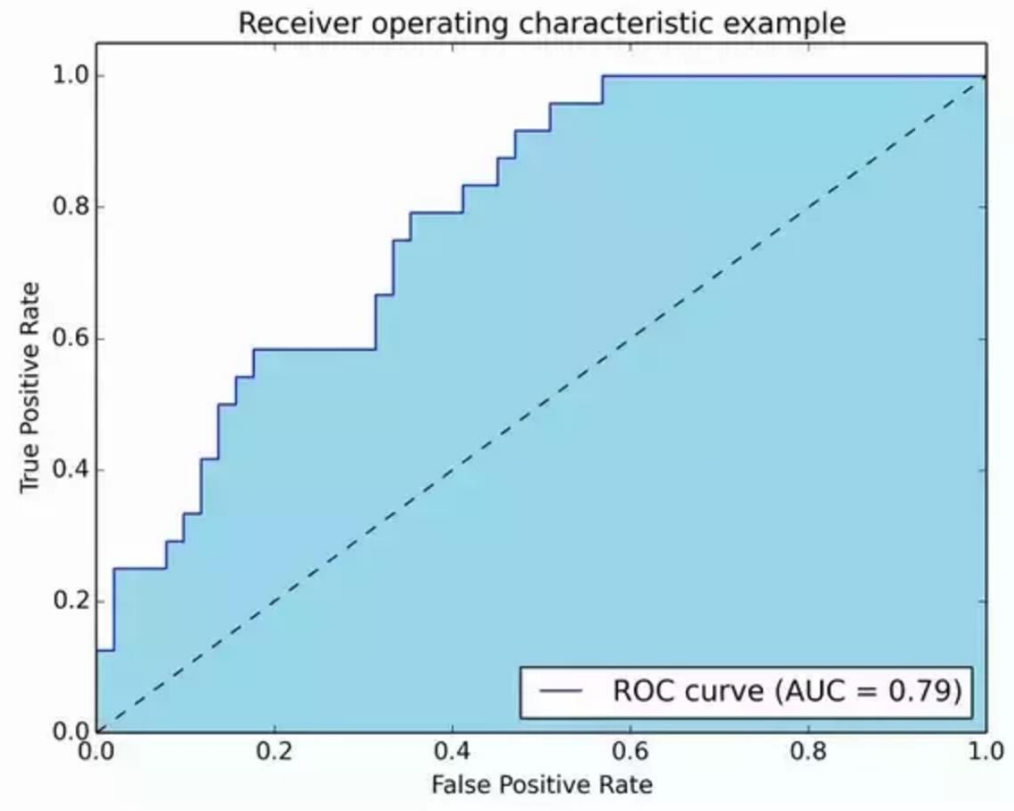
A métrica de log loss penaliza mais classificadores confiantes com uma resposta incorreta.

Logo se otimizarmos a métrica de log loss, tentamos alta confiança nas probabilidade a não ser quando o classificador tenha absoluta certeza da resposta.

AUC

- AUC e GINI são medidas que tentam medir a eficácia do classificador ordenar corretamente as observações e não necessariamente classificar.
- AUC (area under the curve → ROC CURVE) representa a área sob a curva ROC.
- A curva ROC é um gráfico do false positive rate (eixo x) com o true positive rate (eixo y) variando o ponto de corte.
- Um modelo aleatório apresenta uma curva na diagonal e por consequência um AUC de 0.5
- Um modelo perfeito com true positive rate igual a 1 e false positive rate igual a 0, apresenta AUC de 1.

AUC


Receiver operating characteristic example
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.79)

GINI

- Assim como o AUC, só é sensível a ordem das previsões.
- Na verdade, há uma relação um para um com o AUC, em que:

$$GINI = 2*AUC-1$$

Laboratório

Gastos Cartão

Base simulada com 150 observações e 5 variáveis.

- Gastos no cartão em reais
- Idade
- Renda
- Pagamento de impostos
- Segmento

Objetivo:
Ajustar um modelo linear e calcular as métricas de performance estudadas

```
> head(dados)
  Gastos_Cartao Idade Renda Impostos Segmento
1           510    35  1120       60        C
2           490    30  1120       60        C
3           470    32  1040       60        C
4           460    31  1200       60        C
5           500    36  1120       60        C
6           540    39  1360      120        C
```

Exemplos:

Titanic

Prever a probabilidade de sobrevivência dos passageiros do Titanic e calcular as métricas.

Knowledge • 3,464 teams

Titanic: Machine Learning from Disaster

Fri 28 Sep 2012 — Sat 31 De

Dashboard

Home

Data

Make a submission

Competition Details » Get the Data » Make a submission

Predict survival on the Titanic using

Titanic

```
VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown; S = Southampton)
```

Métricas

Exercícios: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

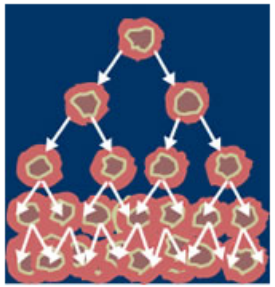
UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Breast Cancer Wisconsin (Diagnostic) Data Set

Download: Data Folder, Data Set Description

Abstract: Diagnostic Wisconsin Breast Cancer Database



Data Set Characteristics:	Multivariate	**Number of Instances:**	569	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	32	**Date Donated**	1995-11-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	390512

Métricas

Base de dados (“cancer.data”) com 699 observações e 10 variáveis de pacientes com tumores. O objetivo é detectar com base em algumas informações dos tumores se é benigno ou maligno.

Remova os dados missing por meio do comando “na.omit”. Poderíamos fazer algo melhor?

Variáveis:

- 1. Sample code number id number
- 2. Clump Thickness 1 - 10
- 3. Uniformity of Cell Size 1 - 10
- 4. Uniformity of Cell Shape 1 – 10
- 5. Marginal Adhesion 1 - 10
- 6. Single Epithelial Cell Size 1 - 10
- 7. Bare Nuclei 1 - 10
- 8. Bland Chromatin 1 - 10
- 9. Normal Nucleoli 1 - 10
- 10. Mitoses 1 - 10
- 11. Class: (2 for benign, 4 for malignant)

Métricas

1.) Utilize seed de 42 e crie amostras de treino (70%) e teste (30%).

2.) Ajuste uma regressão logística.

3.) Como o ajuste se comporta na base de teste? Calcule algumas métricas. Qual você utilizaria.

Seleção de variáveis

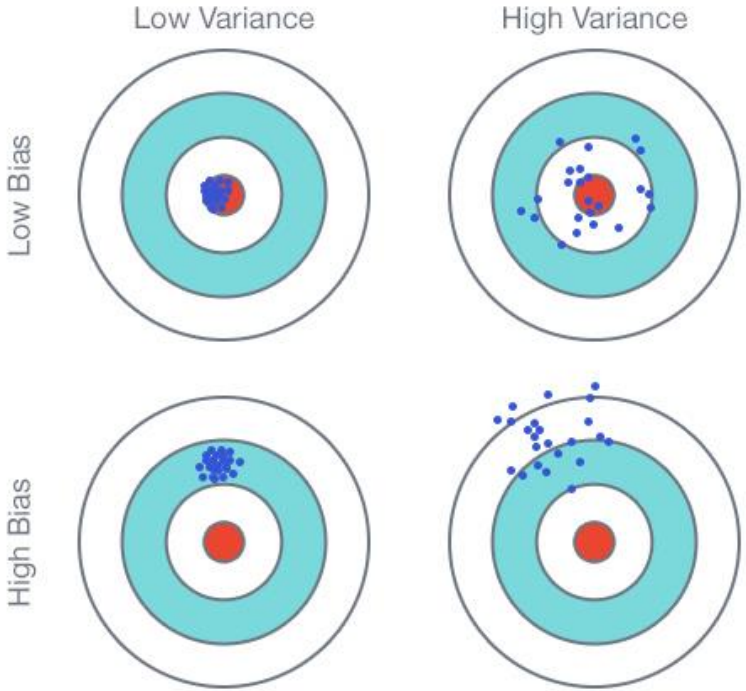
Seleção de variáveis

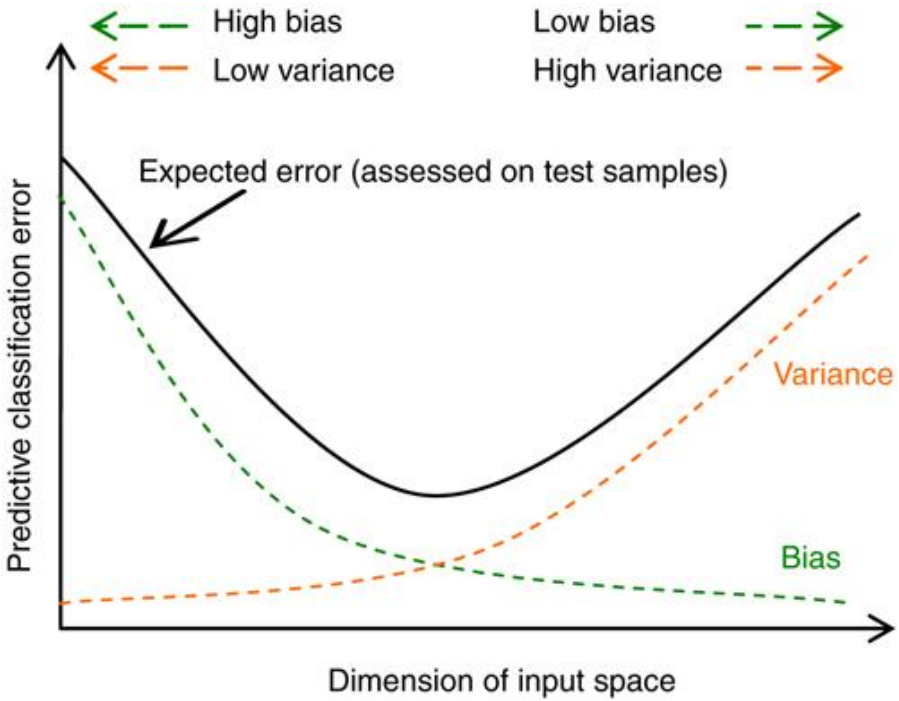
Usar sempre todas as variáveis que temos a disposição é sempre o melhor a se fazer na regressão liner ou na regressão logística?

Quase sempre não! Mas por que?

- **Performance:** os modelos lineares, em geral, apresentam pouco viés, mas alta variância. Assim, forçando alguns parâmetros para 0 pode aumentar a performance do modelo, assim sacrificamos um pouco do viés, mas reduzimos a variância.
- **Interpretação:** com muitas variáveis a interpretação fica muito mais complicada.

Seleção de variáveis





Seleção de variáveis – Best Subset

Uma abordagem para realizar a seleção de variáveis é a chamada best subset. Esta estretégia consiste em construir modelos com todas as combinações possíveis de variáveis e escolher aquele com melhor performance em uma base de dados.

Se temos 10 variáveis, o número total de combinações é 2^10 = 1024 modelos, o que é perfeitamente possível nos dias atuais. Com 20, teremos cerca de um milhão de modelos, o que já começa a ser tornar mais inviável.

Imagine agora iniciando com 1000 variáveis! Teremos que construir 2^1000 modelos, número maior que o número de estrelas no universo.

Seleção de variáveis – Forward

O algoritmo de forward embora não garanta o melhor conjunto de variáveis, tende a fazer uma boa seleção em um tempo bem mais viável.

Em algumas situações, o forward pode até mesmo ter uma melhor performance preditiva, pois apesar de apresentar um viés maior, pode reduzir a variância.

- Inicia-se apenas com o intercepto e a cada passo, adiciona a variável que mais adiciona performance preditiva. O algoritmo termina quando nenhuma variável nova adiciona performance preditiva.

Qual base utilizamos para medir a performance?

- Se usarmos a base de treino para fazer o forward, sempre ao acrescentarmos mais variáveis, o modelo ficará melhor, assim chegaríamos sempre no modelo com todas as variáveis.

- Se usarmos a base de teste, iremos viesar esta base e não teremos mais uma boa estimativa de como o modelo se irá comportar na realidade.

- A estratégia mais comum é particionar os dados em três base de dados (treino, validação e teste). Usamos a base de validação para saber qual seleção de variáveis escolher. Veremos outras técnicas de validação!!!

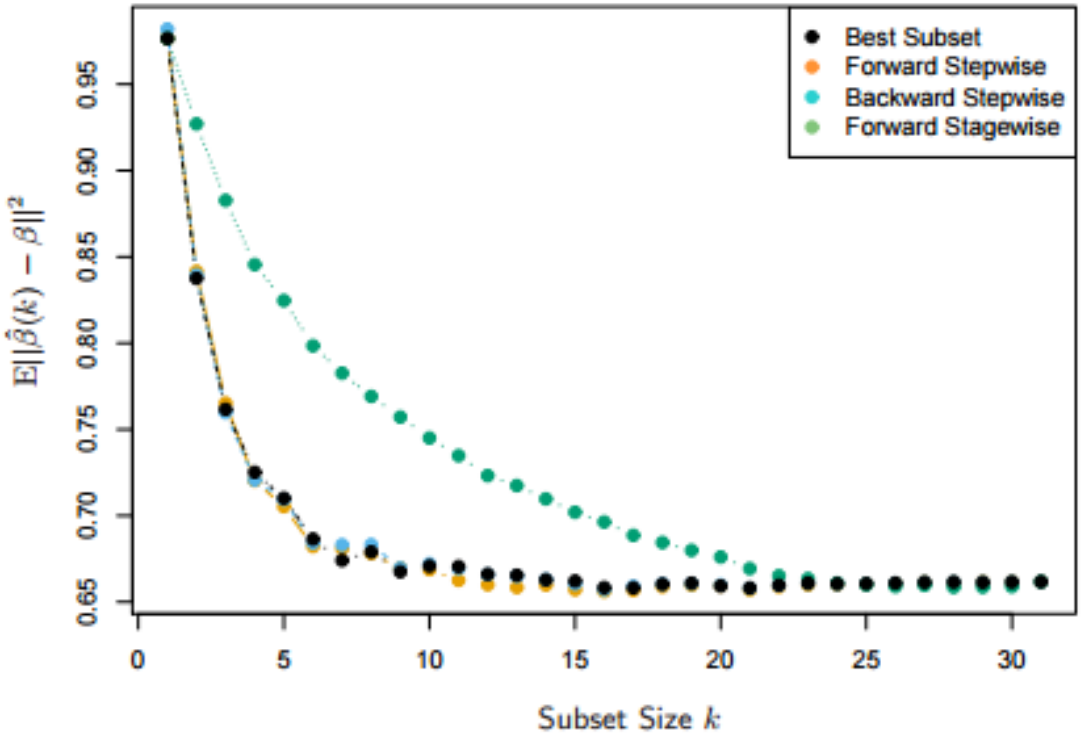
Seleção de variáveis – Backward

- Inicia-se com o modelo completo e retira a cada iteração a variável que menos contribui ao ajuste. Pode se utilizar a performance na base de validação para escolher o ponto de parada.

Seleção de variáveis – Stepwise

- Consiste em utilizar os dois procedimentos (forward e backward) ao mesmo tempo.
- A medida que adicionamos variáveis, em cada etapa, testamos também a fase de backward, que tenta eliminar as variáveis que foram selecionadas anteriormente.
- Método muito popular em Estatística (com diferentes versões, como, por exemplo, olhando a significância da variável).

Seleção de variáveis – Stepwise



Shrinkage Methods - Ridge

- Regressão Ridge ao invés de selecionar as variáveis, limita os valores ajustados dos parâmetros impondo uma penalidade para valores muito grande.

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \right\}.$$

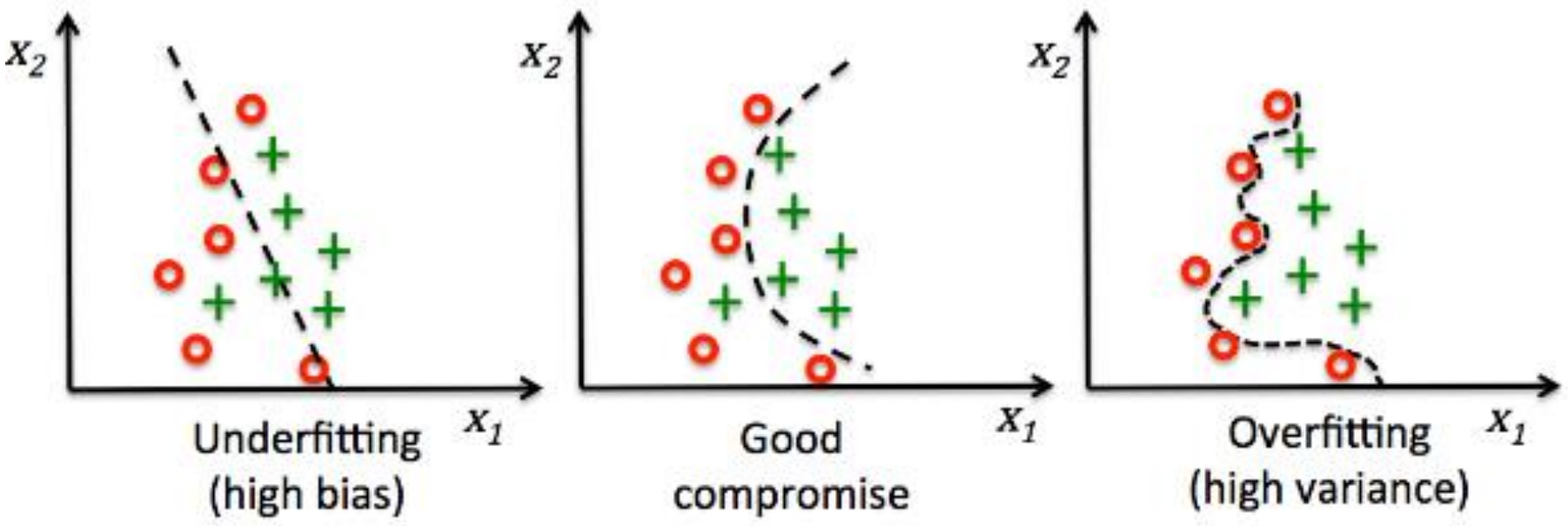
- λ é um parâmetro que controla o grau de penalização. λ muito grande força todos os valores estimados para 0, enquanto λ igual a 0 indica que não temos nenhuma penalização.

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j\right)^2,$$

- Podemos ver a penalização como:

$$\text{subject to} \sum_{j=1}^{P} \beta_j^2 \leq t,$$

Shrinkage Methods - Ridge



Shrinkage Methods - Lasso

- Lasso é muito parecido com o Ridge, com diferenças sutis e importantes:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

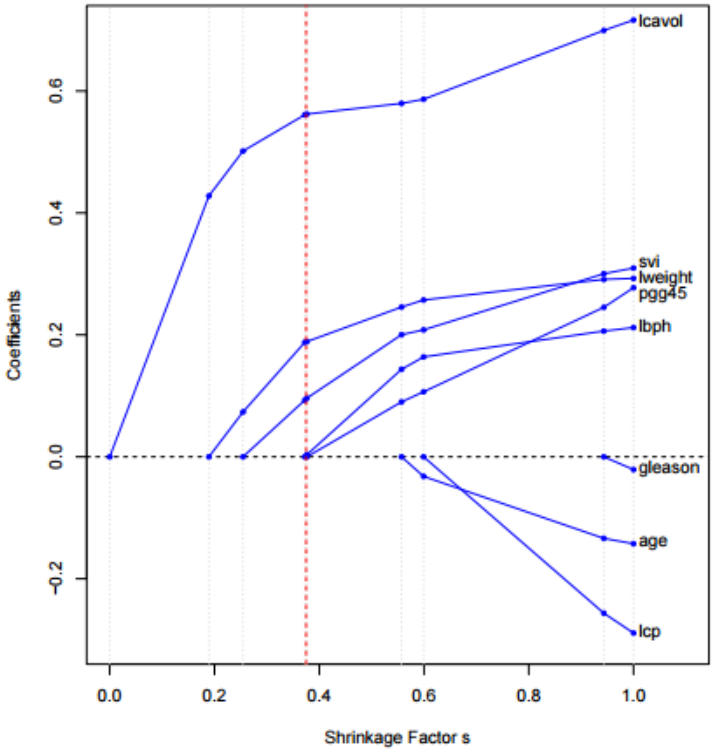
$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Shrinkage Methods - Lasso

- A mudança em relação ao Ridge é a função de penalização, alterando o quadrado pela módulo. Isto traz várias diferenças:
  - Otimização se torna mais problemática. No entanto, a maioria dos softwares apresentam esta opção.
  - Pode estimar coeficientes iguais a 0, o que também pode ser visto como um método para selecionar variáveis.

Shrinkage Methods - Lasso



Laboratório

- Amostra dos dados de Census 1994 dos Estados Unidos.
- O objetivo é prever se o indivíduo recebe mais de 50 mil doláres por ano.
- Variáveis da pessoa que deu entrevista:
  Idade, Tipo de trabalho, educação,
  educação como contínua,
  estado civil, ocupação,
  Posição na família,
  raça, sexo, ganho de capital, perda de capital,
  horas de trabalho por semana, pais de origem.
  http://archive.ics.uci.edu/ml/datasets/Adult

Exercício

- Wine Quality dataset. Base de dados com informações a respeito de vinho verde de Portugal e sua nota de avaliação.

- O objetivo é tentar calcular a probabilidade do vinho ter uma nota alta de avaliação com base nas informações do vinho.

- Variáveis: acidez fixa, acidez volátil,
  ácido cítrico, açúcar residual, calorias,
  dióxido de enxofre livre e total, densidade,
  pH, sulfatos, alcool e qualidade.
  http://archive.ics.uci.edu/ml/datasets/Wine+Quality

Exercício

- Leia e junte as duas bases relativas as variantes dos vinhos (vermelho e branco).
- Crie uma variável binária com o valor 0 caso a nota for inferior a 7 (<) ou 1 caso contrário (>=).
- Crie as amostras de treino e teste (30%) usando seed de 42.
- Execute o stepwise e compare com o modelo completo.
- Divida a base de treino em treino e validação (30%) usando seed de 84.
- Otimize o valor de lambda usando lasso e compare com os ajustes anteriores.
- Faça o mesmo para o Ridge.

Referências Bibliográficas

- Hastie, T., Tibshirani, R. & Friedman, J.H. (2001) “The Elements of Statistical Learning”
- Bishop, C.M. (2007) “Pattern Recognition and Machine Learning”
- Mitchell, T.M. (1997) “Machine Learning”
- Abu-Mostafa, Y., Magdon-Ismail, M., Lin, H.T (2012) “Learning from data”
- Theodoridis, S., Koutroumbas, K., (2008) “Pattern Recognition”
- Kuhn, M., Johnson, K., (2013) “Applied Predictive Modeling”
- Burns, P. (2011) “The R inferno”