Klątwa wielowymiarowości
The Curse of Dimensionality
2. część

Agnieszka Pocha
Michał Kowalik

18 marca 2015

na podstawie książki:
Bertrand Clarke, Ernest Fokoue, Hao Helen Zhang
Principles and Theory for Data Mining and Machine Learning

Agenda

Klątwa wielowymiarowości

Przy wysokim wymiarze przestrzeni, dane są zbyt rzadkie.
Przy wysokim wymiarze przestrzeni, liczba możliwych modeli do rozważenia rośnie zbyt szybko.

In general, the number of polynomial models of order at most 2 in p variables is $2a - 1$, where $a = 1 + 2p + \frac{p(p-1)}{2}$.

Przykład

Dla $p = 1$ jest 7 możliwych różnych modeli:

$\mathbb{E}(Y) = \beta_0,$
$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1,$
$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2,$

$\mathbb{E}(Y) = \beta_1 x_1,$
$\mathbb{E}(Y) = \beta_0 + \beta_2 x_1^2,$

$\mathbb{E}(Y) = \beta_2 x_1^2,$
$\mathbb{E}(Y) = \beta_1 x_1 + \beta_2 x_1^2,$

Principal Component Analysis - Analiza głównych składowych

Celem PCA jest taki obrót układu współrzędnych, aby maksymalizować w pierwszej kolejności wariancję pierwszej współrzędnej, następnie wariancję drugiej współrzędnej, itd..

PCA jest często używana do zmniejszania rozmiaru zbioru danych statystycznych, poprzez odrzucenie ostatnich czynników.

Wariancja

Klasyczna miara zmienności. Intuicyjnie utożsamiana ze zróżnicowaniem zbiorowości.

$$D^2(X) = E(X^2) - [E(X)]^2$$

E - wartość oczekiwana

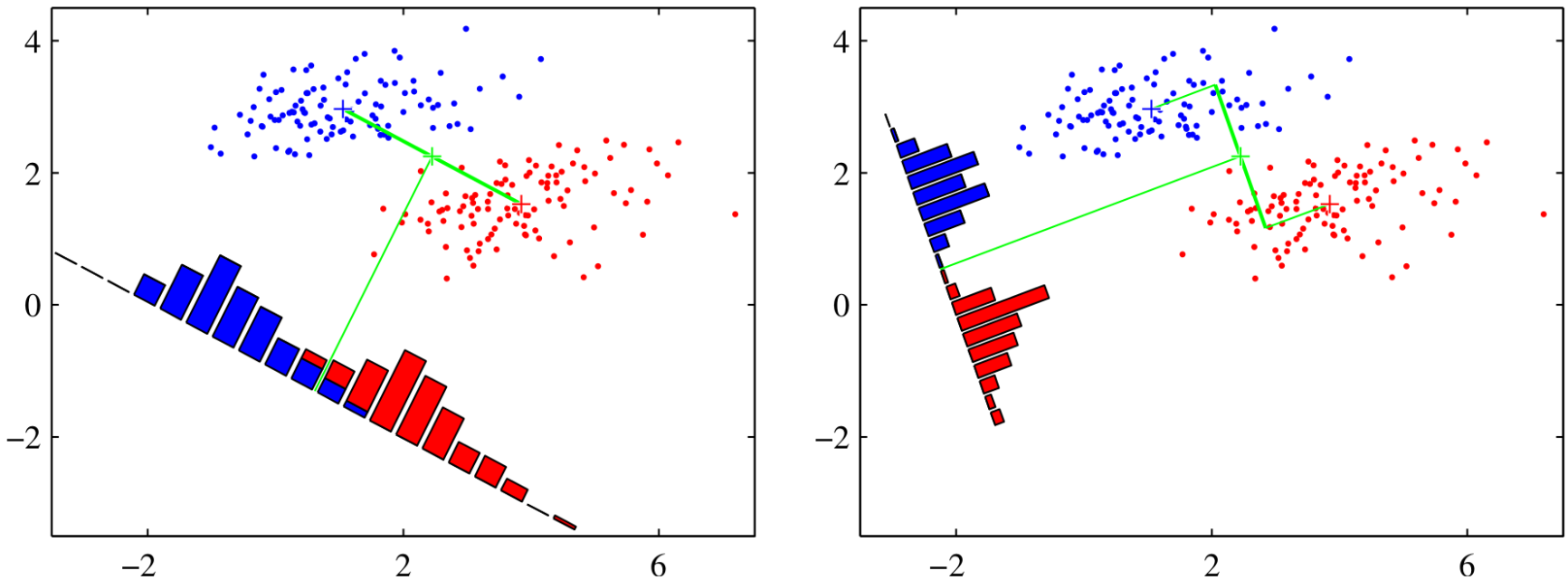
Algorytm

- Obliczenie wartości średniej dla każdej cechy: $u[m] = \frac{1}{N} \sum_{n=1}^{N} X[m,n]$
- Policzenie wartości odchyleń dla każdej komórki danych: $B[i,j] := X'[i,j] = X[i,j] - u[i]$
- Wyznaczenie macierzy kowariancji
- Policzenie wartości własnych macierzy kowariancji

  wartość własna opisujące endomorfizm tj. przekształcenie danej przestrzeni liniowej w samą siebie.
- Wybór wartości własnych: można dokonać zawężenia wymiaru przestrzeni.

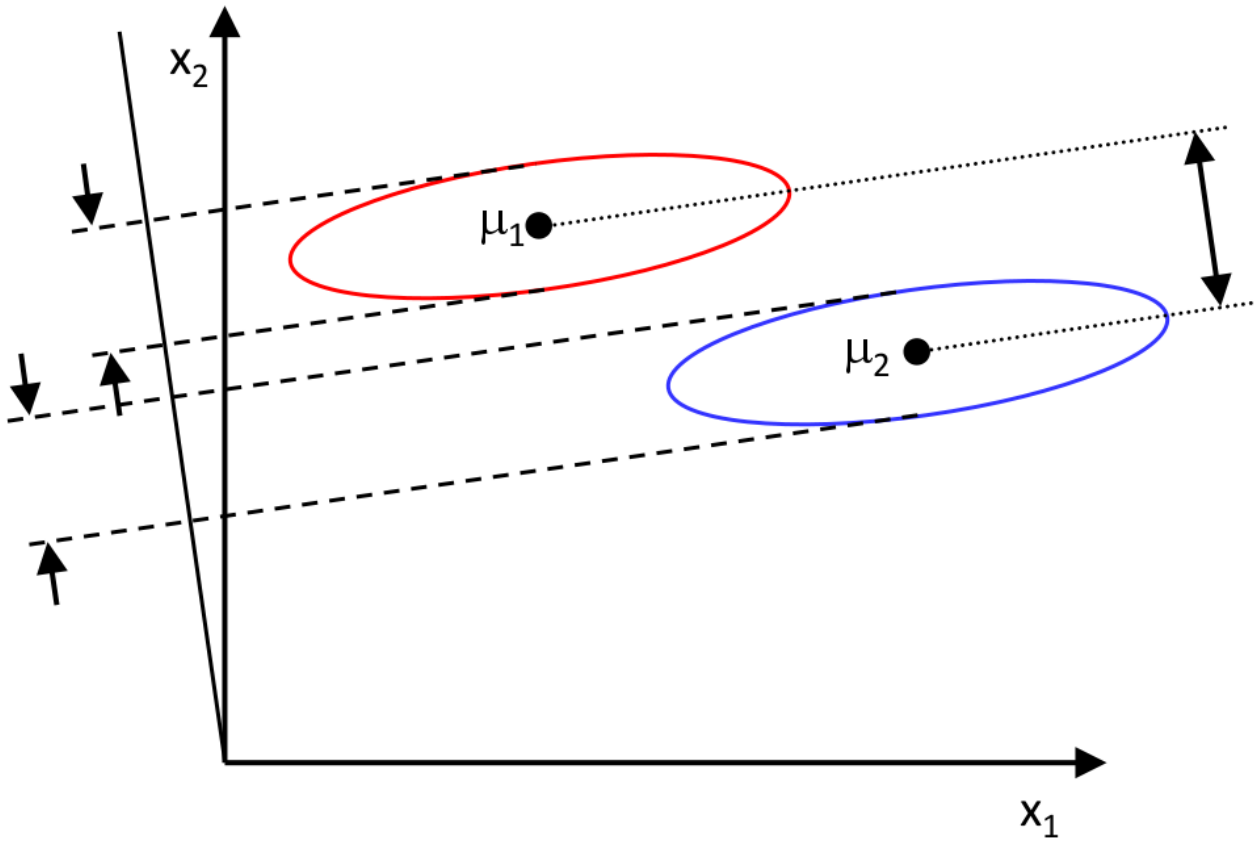
  Im wyższa wartość własna tym odpowiadający jej wektor własny jest słabiej skorelowany z pozostałymi.
- Wyznaczenie wektorów własnych
- Rzutowanie na wektory własne

LDA



Przykład

na modelu: $y = w^T X$


x_2
μ_1
μ_2
x_1

Wybór Modelu

oszacowanie jakości róznych modeli w celu wybrania najlepszego.

Dopasowanie Modelu

po wyborze najlepszego modelu, oszacowanie jego błędu przewidywania dla nowych danych.

Wybór Modelu

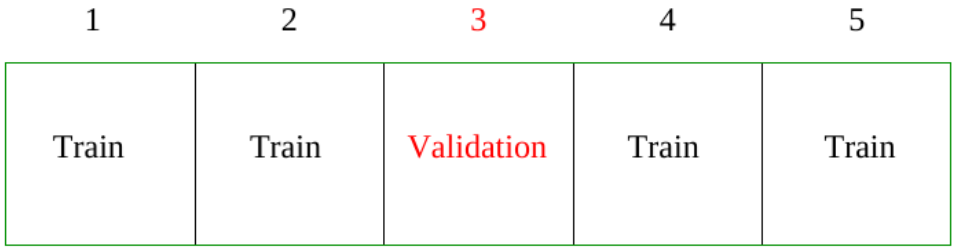
oszacowanie jakości róznych modeli w celu wybrania najlepszego.

Dopasowanie Modelu

po wyborze najlepszego modelu, oszacowanie jego błędu przewidywania dla nowych danych.

Kryteria

- AIC, BIC, Cp - metody analityczne
- walidacja krzyżowa, Bootstrap - wielokrotne używanie punktów w zbiorach danych.



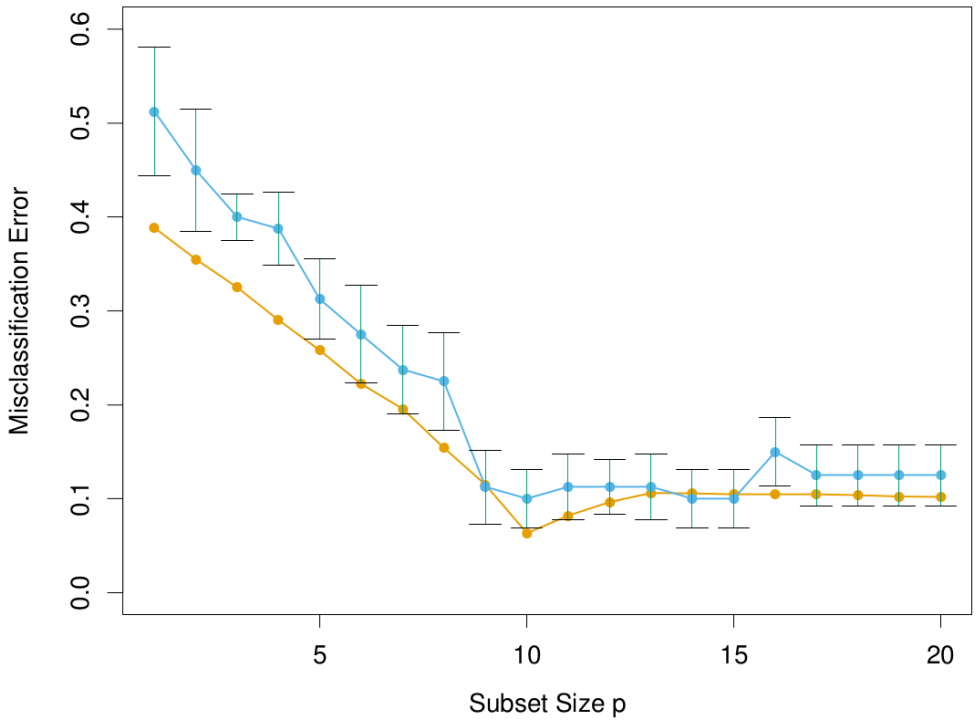
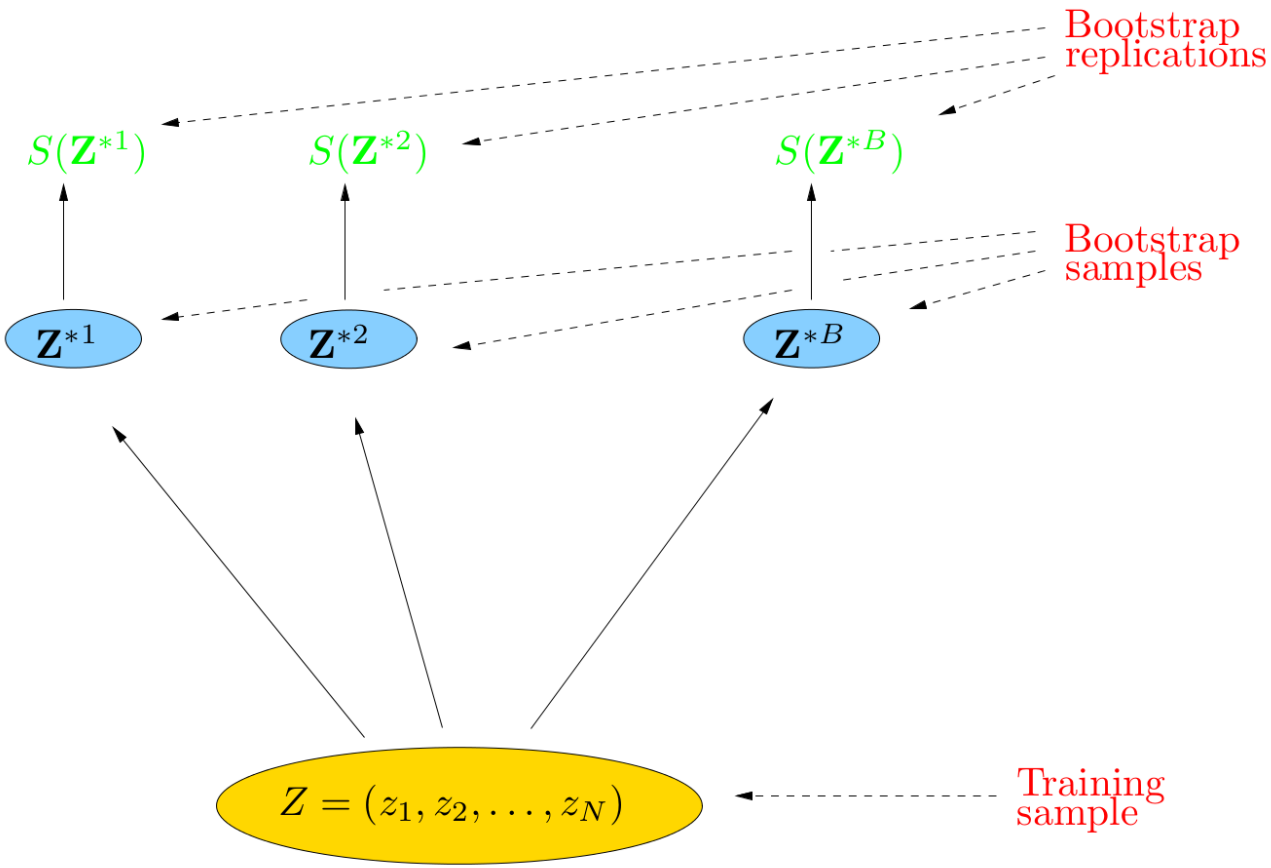


FIGURE 7.9. *Prediction error (orange) and tenfold cross-validation curve (blue) estimated from a single training set, from the scenario in the bottom right panel of Figure 7.3.*

Bootstrap

Bootstrap, to ogólna metoda do oceny statystycznej dokładności.

Główna idea polega na tworzeniu nowych zbiorów poprzez losowanie ze zwracaniem punktów ze zbioru trenującego. Każdy nowy zbiór powinien mieć tę samą wielkość co zbiór trenujący. Tworzy się B nowych zbiorów, następnie dopasowuje się model do każdego z tych nowych zbiorów i sprawdza się jego zachowanie po nauczeniu na każdym z tych nowych zbiorów.



Cel

Oszacowanie dokładności $S(Z)$ policzonej na podstawie naszego zbioru danych.

$S(Z)$

$S(Z)$ - dowolna wartość policzona na podstawie danych Z, np. predykcja na podstawie jakiegoś punktu wejściowego.

Na podstawie 'Bootstrap sampling', można oszacować dowolny aspekt dotyczący rozkładu $S(Z)$.

$$\widehat{Var}[S(Z)] = \frac{1}{B-1}\sum_{b=1}^{B}(S(Z^{*b}) - \tilde{S^*})^2$$

gdzie:

$$(mean)\ \tilde{S^*} = \frac{1}{B}\sum_{b} S(Z^{*b})$$

Jak zaaplikować Bootstrap do oszacowania błędu predykcji?

Jeśli $\hat{f}^{*b}(x_i)$ to oszacowana wartość w x_i na modelu dopasowanym do b-tego nowego zbioru, to szacowana wartość to:

$$\widehat{Err_{boot}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{f}^{*b}(x_i))$$

Jak zaaplikować Bootstrap do oszacowania błędu predykcji?

Jeśli $\hat{f}^{*b}(x_i)$ to oszacowana wartość w x_i na modelu dopasowanym do b-tego nowego zbioru, to szacowana wartość to:

$$\widehat{Err_{boot}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{f}^{*b}(x_i))$$

Lepiej

$$\widehat{Err_{boot}}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

Jak zaaplikować Bootstrap do oszacowania błędu predykcji?

Jeśli $\hat{f}^{*b}(x_i)$ to oszacowana wartość w x_i na modelu dopasowanym do b-tego nowego zbioru, to szacowana wartość to:

$$\widehat{Err_{boot}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{f}^{*b}(x_i))$$

Lepiej

$$\widehat{Err_{boot}}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

Średnia liczba różnych obserwacji w każdym nowym zbiorze, to około $N \cdot 0.632$.

Akaike zaproponował, aby wybierać ten model dla którego najmniejsza jest wartość:

$$\mathrm{AIC} = -2 \sum_j \ln(\hat{\pi}_j) + 2q$$

gdzie:
$\hat{\pi}_j$ – estymowane prawdopodobieństwo, przy założeniach danego modelu, uzyskania takiej właśnie wartości obserwacji j jaka była naprawdę uzyskana.
q – liczba parametrów modelu

Bayesian information criterion (BIC) jest formalnie zdefiniowane jako:

$$\mathrm{BIC} = -2 \cdot \ln \hat{L} + k \cdot \ln(n).$$

gdzie:
n - liczba elementów/przykładów w zbiorze danych
k - Liczba parametrów modelu.
$\hat{L}$ - Zmaksymalizowana wartość funkcji prawdopodobieństwa dla modelu M

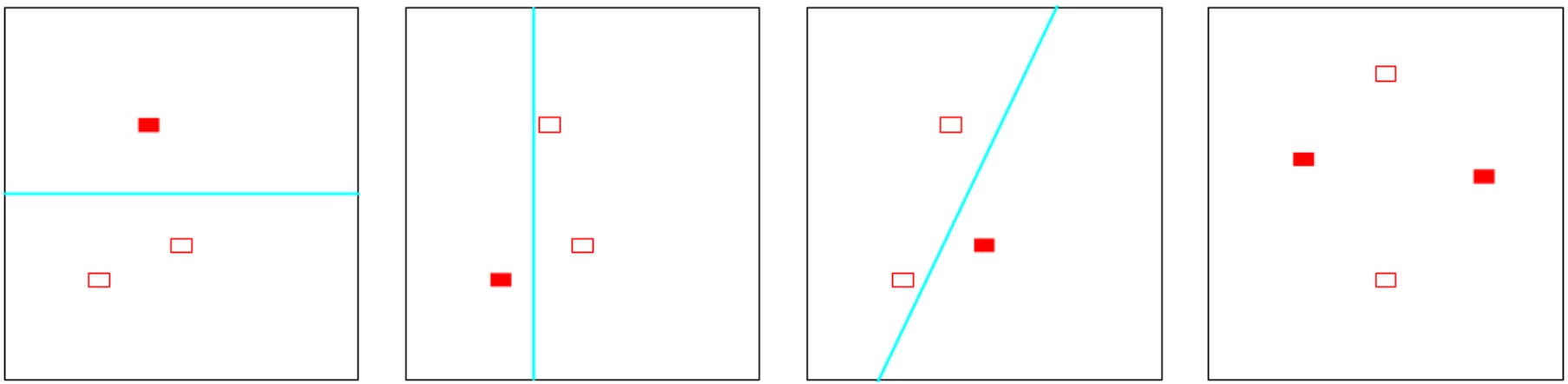
Vapnik–Chervonenkis (VC) dimension

Wymiar VC dla klasy $\{f(x, \alpha)\}$ jest zdefiniowane przez najwiekszą liczbę takich punktów (w dowolnej konfiguracji), że można je sklasyfikować przez elementy z $\{f(x, \alpha)\}$.

Klasyfikowalność zbioru

Zbiór punktów jest klasyfikowalny przez klasę funkcji jesli, nie ważne jak przyporzadkujemy etykiety binarne do każdego z punktów, istnieje taki element klasy, który może je idealnie separować.

Wymiar VC funkcji liniowych



W ogóle, liniowy klasyfikator w przestrzeni p-wymiarowej ma wymiar VC o rozmiarze $p + 1$.
Ponadto, można pokazać, że rodzina $\sin(\alpha x)$ ma niekończony wymiar VC.

Często minimalizacja AIC, CV lub Bootstrap wskazuje na model dosyć bliski najlepszego.
Jednakże AIC bywa niepraktyczny.