An introduction to Statistical Learning

Lila Gmerek / Week 1 / 29.04.2019

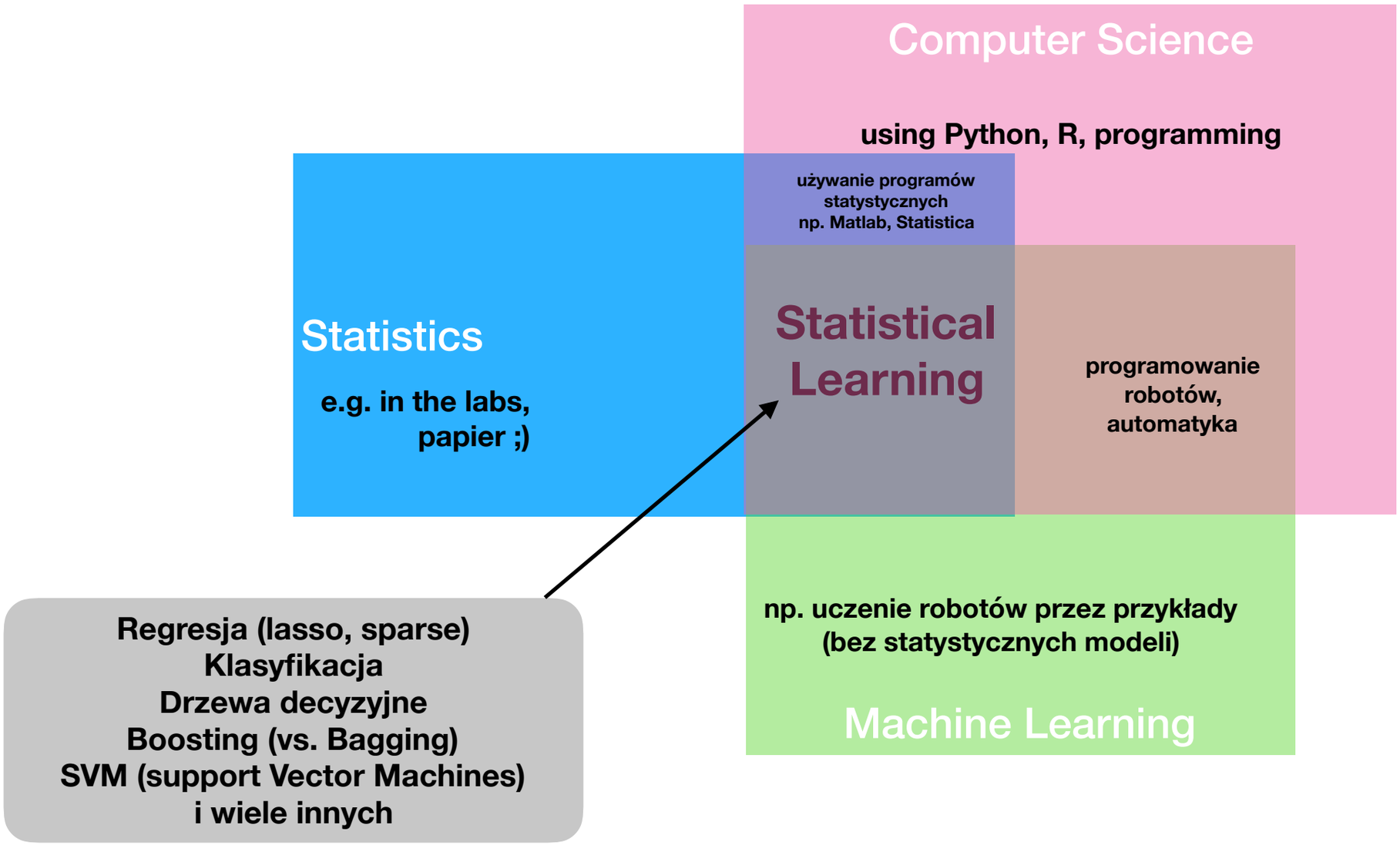
Links

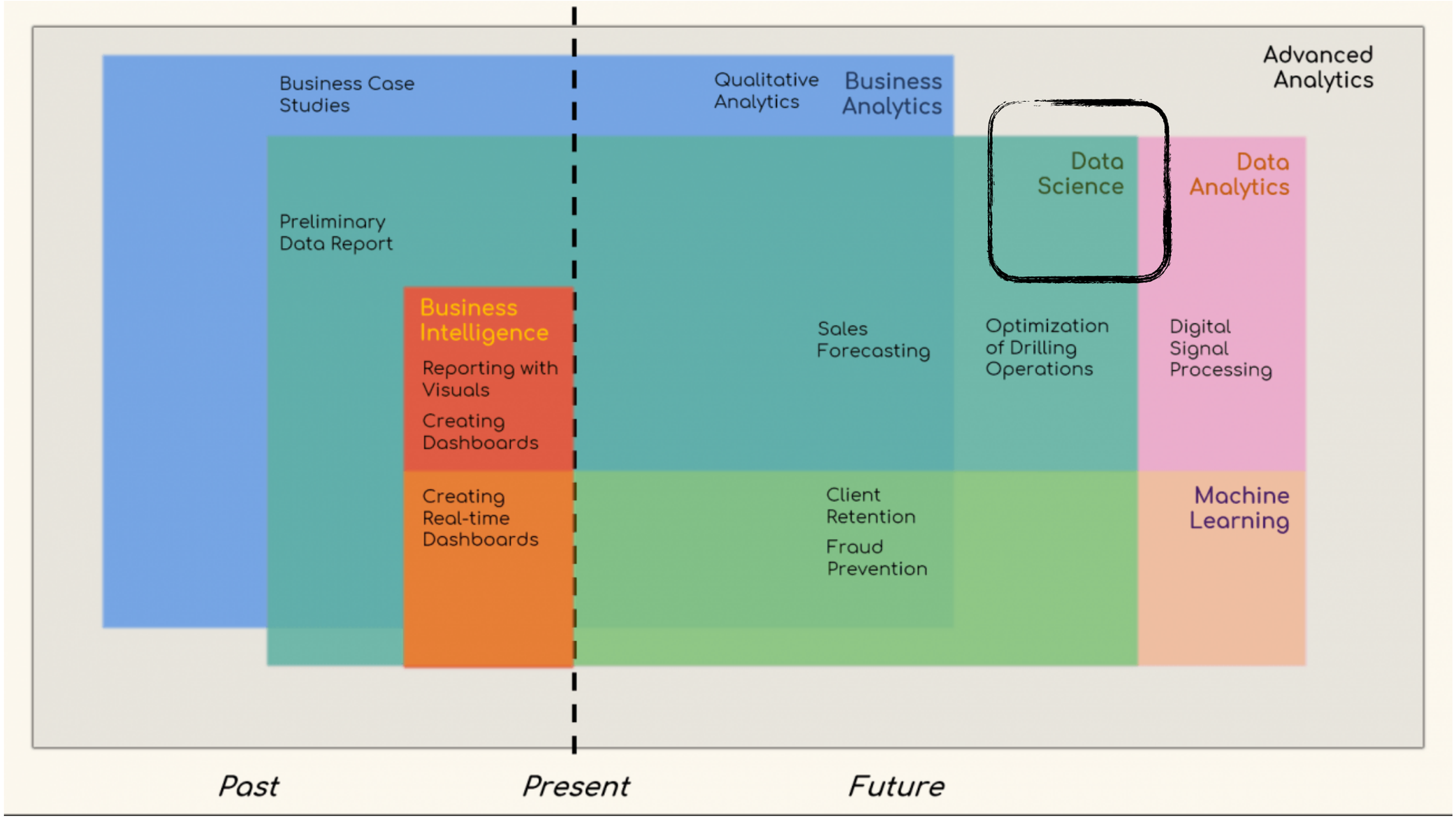
- Book http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf

- Github: https://github.com/erg0-0/ML-study-group-pl-fc

Introduction. An Overview of Statistical Learning

Statistical learning refers to a set of tools for modeling and understanding complex datasets. It is a recently developed area in statistics and blends with parallel developments in computer science and, in particular, machine learning. The field encompasses many methods such as the lasso and sparse regression, classification and regression trees, and boosting and support vector machines.





https://sintelix.com/blog/machine-learning-v-data-science-v-analytics/

Statistical learning refers to a vast set of tools for understanding data. These tools can be classified as supervised or unsupervised. Broadly speaking, supervised statistical learning involves building a statistical model for pre- dicting, or estimating, an output based on one or more inputs. Problems of this nature occur in fields as diverse as business, medicine, astrophysics, and public policy. With unsupervised statistical learning, there are inputs but no supervising output; nevertheless we can learn relationships and struc- ture from such data.

Narzędzia:

- uczenie nadzorowane
- uczenie nienadzorowane (uczenie zależności, struktur)

CEL:

- stworzenie modelu predykcyjnego/szacującego
- rezultat (*output*) wynika z danych wejściowych (1+ *input*)

Dyskusja:

- Jaka jest różnica między Machine Learningiem a Data Science?

- Czy Machine Learning to Deep learning czy Deep learning to machine learning?

- Co ma wspólnego big data (engineering) z machine learningiem?

- Deep learning WIkipedia: https://en.wikipedia.org/wiki/Deep_learning
- Machine Learning / Uczenie maszynowe: https://pl.wikipedia.org/wiki/Uczenie_maszynowe
- Data Science: https://en.wikipedia.org/wiki/Data_science
- Big Data: https://en.wikipedia.org/wiki/Big_data

Przykłady

1. WAGE DATA - analiza danych na temat zarobków

Cel analizy: analiza czynnikow, ktore wplywaja na wysokosc zarobków (wage) grup mężczyzn z regionu atlantyckiego w US.

Featery/cechy analizowane / input : wiek (age), edukacja (education), rok(year) —> wplyw na zarobki (wage)

Natura problemu: regresja

Oczekiwany output: (procentowy) wpływ danego czynnika na zarobki / prawdopodobieństwo etc.

UWAGA! znamy te dane dla konkretnej grupy juz! Chcemy to przewidywać w przyszłosci też dla innych grup!

Wnioski na bazie wykresów:

Czy lepiej analizować te czynniki oddzielnie, czy razem?

1. Wage jako funkcja wieku. Zarobki rosną wraz z wiekiem, ale spadaja po 60ym roku zycia (fig. 1, niebieska linia)

2. Wage jako funkcja roku. Wzrost o ok. 10k$ miedzy 2003-2009

3. Wage jako funkcja edukacji. Zarobki rosna wraz z poziomem edukacji.

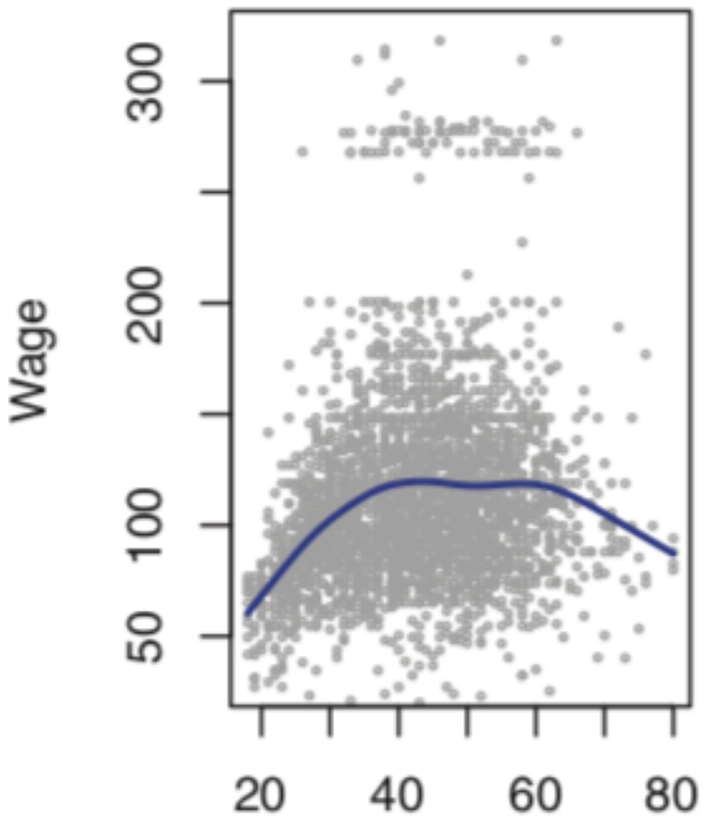
Odpowiedz: Najlepszą predykcję da kombinacja tych 3 czynników: age, education, year. Analizowanie ich oddzielnie może zaburzyć predykcję.

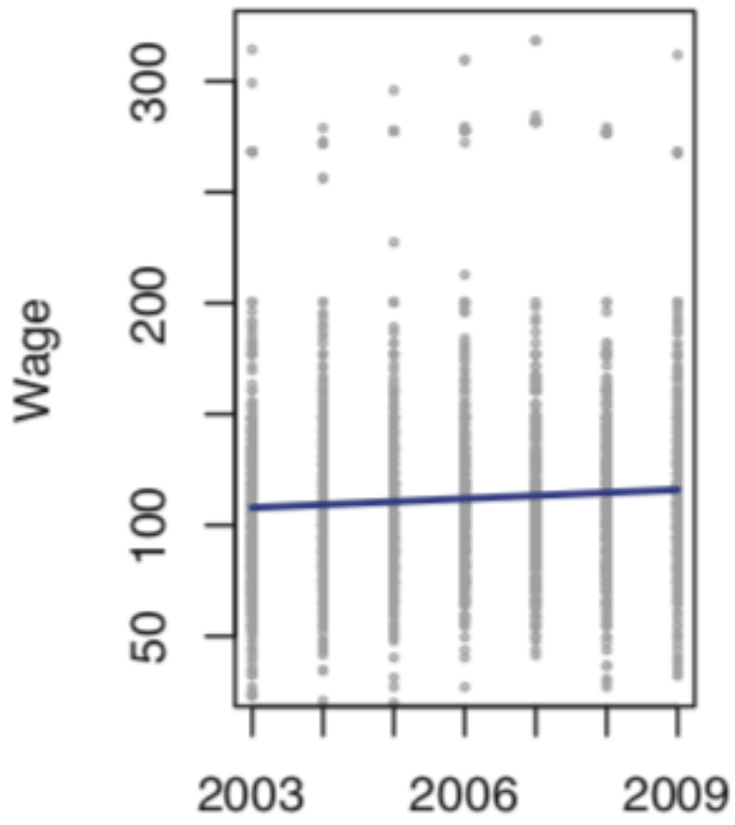
Jakiego modelu predykcyjnego użyc?

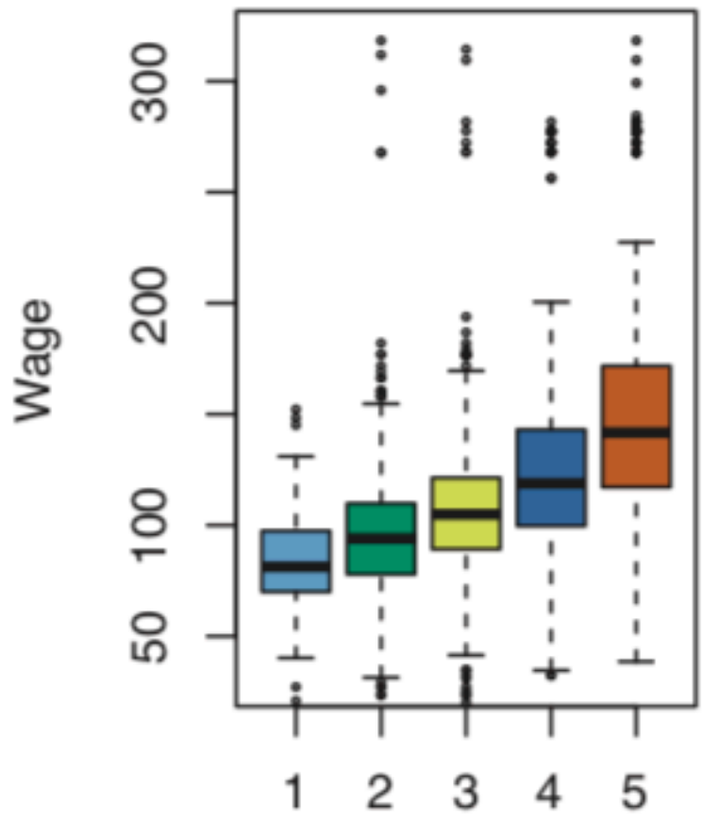
Supervised learning - uczenie maszynowe nadzorowane

1. Rozdzial 1 - regresja liniowa (widac zwlaszcza na wykresie year/wage)

2. Rozdzial 7 - modelowanie nie-liniowe, regresja wielomianowa , GAMs itp.


Wage
50
100
200
300
20
40
60
80
Age


Wage
50
100
200
300
2003
2006
2009
Year


Wage
50
100
200
300
1
2
3
4
5
Education Level

Kiedy problem zarobków jest natury regresyjnej?

The Wage data involves predicting a continuous or quantitative output value.

Wartości predykowane są liczbami ciągłymi, a zasadniczo po prostu mają wartość numeryczną, ciągłą (tzw. float).

- jaka jest prawdopodobieństwo, że polski klient kupi nowe płatki naszej firmy ‚WasabiFlakes' ?
- jaka jutro będzie temperatura powietrza? jaka będzie cena nieruchomości w Bostonie?
- o ile wzrośnie/spadnie sprzedaż płatków Wasabi Flakes w następstwie celowej kampanii marketingowej?

Kiedy ten problem jest natury klasyfikacyjnej?

Jeśli chcesz przewidywać wartości nie-numeryczne - kategoryczne albo jakościowe (categorical or qualitative)

- czy klient kupi WasabiFlakes?(tak/nie) , czy pacjent będzie miał raka? (tak/nie)
- czy jutro będzie padać? (tak/nie); czy to jest kwiat irys setosa, versicolor czy virginica? (clustering)
- do których klientów zaadresować kampanię płatków WasabiFlakes? [klastryzacja, grupowanie]

2. Stock Market Data (Smarket Data)

Cel predykcji: *The goal is to predict whether the index will increase or decrease on a given day using the past 5 days' percentage changes in the index.* (nie przewiduje się wartości numerycznej, ale kategoryczna!)

Natura problemu: klasyfikacja

Featery/cechy analizowane/ input: indeks giełdowy, data (dzień)

Oczekiwany output: czy market performance(obserwacja) idzie w górę <Up bucket> czy w dół <Down bucket>

Uwaga! Mamy dane historyczne! Chcemy to przewidzieć w przyszłości!

Wnioski na bazie wykresów:

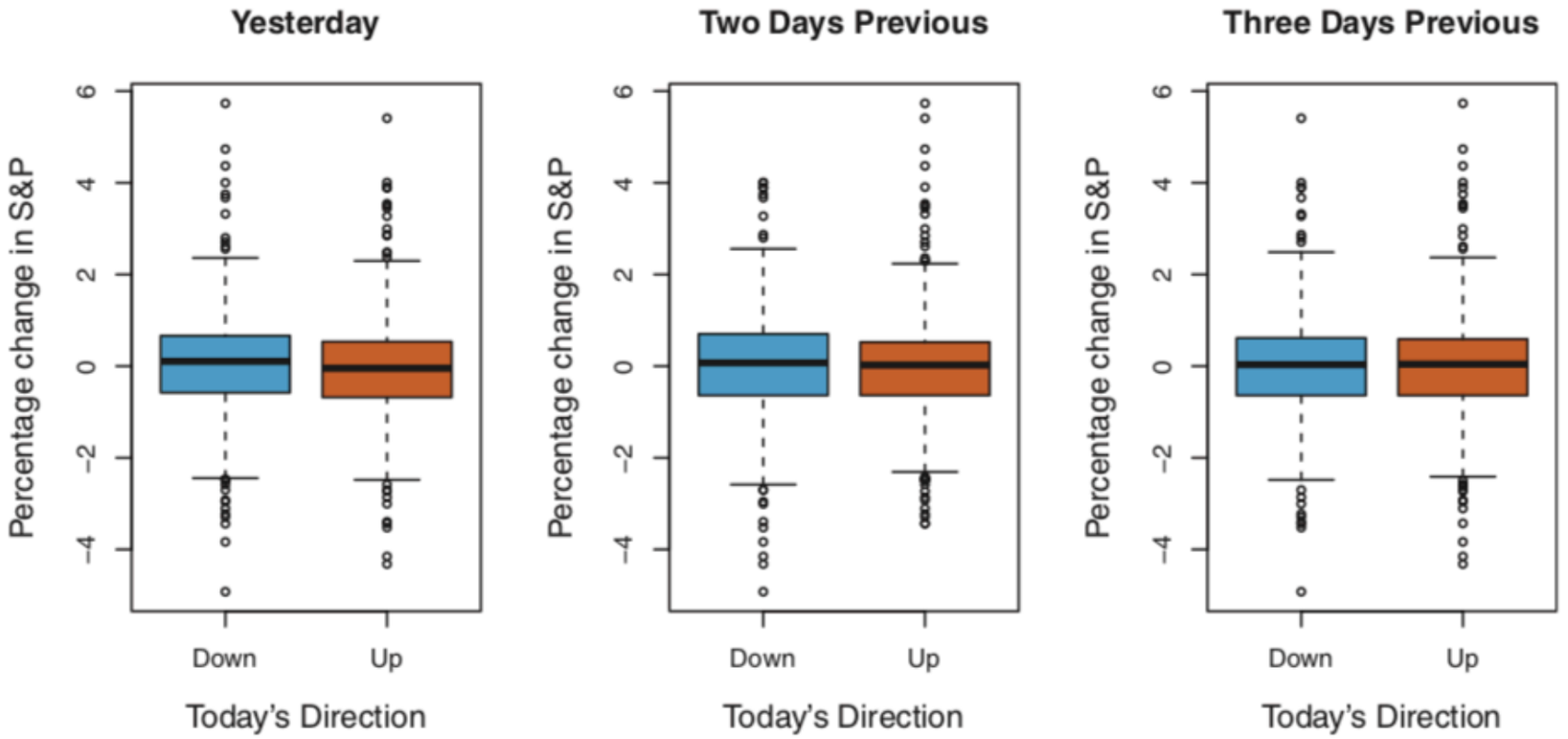
1. Nie ma prostej zależności między poprzednim dniem(dniami) a wysokością indeksu.

Gdyby była, gra na giełdzie nie byłaby taka trudna, a wszyscy byliby bogaci ;-).

2. Są modele, które oceniają z 60% dokładnością trend indeksu w cyklu 5-letnim.

Jakiego modelu predykcyjnego użyc?

Rozdział 4 - metody statystyczne klasyfikacyjne - Supervised Learning


Yesterday
Two Days Previous
Three Days Previous
Percentage change in S&P
6
4
2
0
−2
−4
Down
Up
Today's Direction

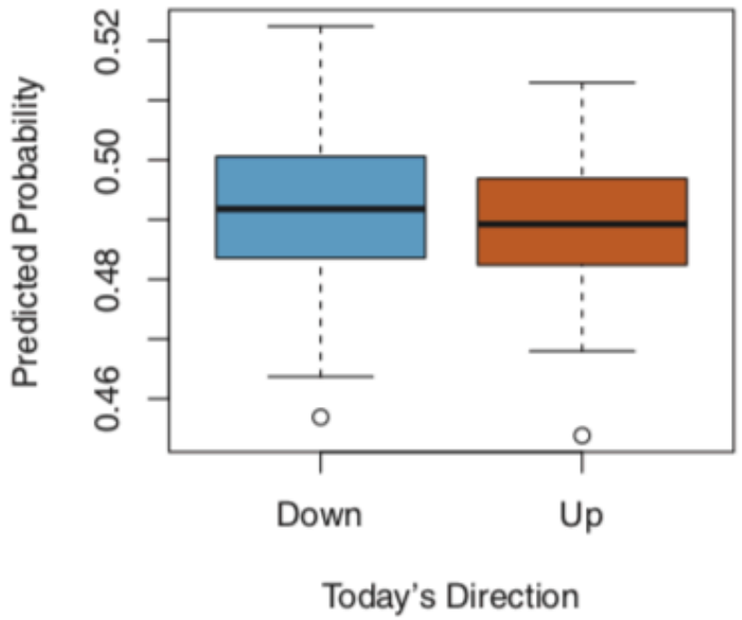


FIGURE 1.3. *We fit a quadratic discriminant analysis model to the subset of the* `Smarket` *data corresponding to the 2001–2004 time period, and predicted the probability of a stock market decrease using the 2005 data. On average, the predicted probability of decrease is higher for the days in which the market does decrease. Based on these results, we are able to correctly predict the direction of movement in the market 60% of the time.*

3. Gene Expression Data - NCI60

Cel predykcji:

In a marketing setting, we might have demographic information for a number of current or potential customers. We may wish to understand ***which types of customers are similar to each other by grouping individuals according to their observed characteristics.***

Dataset NCI60 - we are interested in determining whether there are groups, or clusters, among the cell lines based on their gene expression measurements

Natura problemu: (klasyfikacja) **clustering - NIE predykcja, analiza teraźniejszości, NIE przyszłości**

Featery/cechy analizowane/ input: dane o klientach (X cech m.in. historia zakupów itd.)

Dataset NCI60 - 6.830 obserwacji 64 linii komórek rakowych(? cancer cell lines)

Oczekiwany output: algorytm ma pogrupować klientów na podstawie najlepszych znalezionych zależności

Algorytm ma pogrupować cell lines na podstawie ich pomiarów ekspresji genetycznej. Nie wiemy jak - gdybyśmy wiedzieli - nie trzebaby tworzyć modelu. Danych jest zbyt dużo!

Uwaga! NIE MA wartości oczekiwanej! Nie wiemy jak pogrupować klientów - algorytm sam ma na to „wpaść”!

Wnioski na bazie wykresów NCI60:

1. Aby przedstawić dane na wykresie zastosowano redukcję wymiarów (dimentions) do dwóch (fig. 1.4), co umożliwiło podział na 4 główne grupy. Każdy punkt to jeden z 64 cell lines

2. Drugi wykres to pierwszy + uwzględnienie 14 rodzajów raka za pomocą koloru+kształtu. Widać pewne zależności np. konkretny rodzaj raka pojawia się blisko siebie w danym klastrze linii komórkowych.

3. Utrata informacji jest nieunikniona.

Jakiego modelu użyc?

Rozdział 10 - metody statystyczne klasyfikacyjne - Uczenie Maszynowe nienadzorowane (Unsupervised Learning)

Brief History of Statistical Learning

początek XIX w.

Legendre & Gauss publikują *method of least square (metodę najmniejszego kwadratu), w której zastosowano pierwszą formę regresji liniowej dziecina: Astronomia*

1936

Fisher wprowadza „linear discriminant analysis” co ma pomóc w problemach natury klasyfikacyjne j

1940

różni autorzy proponują alternatywe dla problemów klasyfikacji: **regresję logistyczną!**

lata 70.

Nelder i Wedderburn tworzą pojęcie „ generalized linear models” na określenie całej klasy metod uczenia statystycznego, do których zaliczają się regresja logistyczna i liniowa jako „special cases”

początek 80.

nareszcie komputery mają taką moc obliczeniową, że można projektować „nieliniowe modele predykcyjne”

poł 80.

Breiman, Friedman, Olshen & Stone wprowadzają drzewa klasyfikacyjne i regresyjne, są wśród pierszych, którzy pokazują efektywność praktycznej i szczegółowej implementacji metody łacznie z zastosowaniem „cross-validation” („sprawdzianu krzyżowego” do wyboru właściwego modelu.

1986

Hastie & Tibshirani stworzyli pojęcie”generaliz ed additive models „ na określenie klasy nielinearnych metod uzupełniających do istniejących już „generalized linear models”. Stworzyli również praktyczną implementację software’ową.

Od tego czasu Statistical Learning, a potem i bardziej modne Data Science przeżywa rozkwit. Dostępność software’u oraz mocy obliczeniowych, a także przyjazność dla użytkownika nie stanowią już takiej przeszkody dla szerszej społeczności.