# Uczenie z nadzorem - wprowadzenie

Marcin Sidorowicz

Seminarium magisterskie 2, spec. Analiza danych

10 marca 2021

# Rozumienie intuicyjne

W uczeniu z nadzorem analizujemy (lub mierzymy) dane wejściowe, które mają wpływ na dane wyjsciowe.

# Rozumienie intuicyjne

W uczeniu z nadzorem analizujemy (lub mierzymy) dane wejściowe, które mają wpływ na dane wyjsciowe.
Zmienne wejściowe często nazywa się *predyktorami* lub zmiennymi objaśniającymi, natomiast zmienne wyjściowe to *odpowiedzi* lub zmienne zależne.

# Rodzaje zmiennych

Wejście i wyjście zależą od analizowanego zestawu danych. Mogą być one *ilościowe*, tzn. miedzy pomiarami można wprowadzić relację "$A > B$" bądź podobną, lub jakościowe - tzn. przyjmują wartość z pewnego zbioru (np. dla zestawu *iris* są to *virginica*, *setosa*, *versicolor*.

# Rodzaje zmiennych

Wejście i wyjście zależą od analizowanego zestawu danych. Mogą być one *ilościowe*, tzn. miedzy pomiarami można wprowadzić relację "$A > B$" bądź podobną, lub jakościowe - tzn. przyjmują wartość z pewnego zbioru (np. dla zestawu *iris* są to *virginica*, *setosa*, *versicolor*.
Zmienne jakościowe nazywane są też dyskretnymi lub kategorycznymi.

# Rodzaje zmiennych

Wejście i wyjście zależą od analizowanego zestawu danych. Mogą być one *ilościowe*, tzn. miedzy pomiarami można wprowadzić relację "$A > B$" bądź podobną, lub jakościowe - tzn. przyjmują wartość z pewnego zbioru (np. dla zestawu *iris* są to *virginica, setosa, versicolor*.
Zmienne jakościowe nazywane są też dyskretnymi lub kategorycznymi.
Zmienną o typie pośrednim jest zmienna o typie ordynalnym, tzn. skończony zbiór, na którym wprowadzamy intuicyjny porządek bez metryki (np. *mały, średni, duży*).

# Reprezentacja zmiennych

Zmienne jakościowe mogą być przedstawiane liczbowo. W najprostszym przypadku (binarnym) wystarczy zapisać je jako 0/1, albo 1/-1.

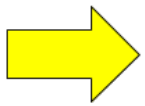
# Reprezentacja zmiennych

Zmienne jakościowe mogą być przedstawiane liczbowo. W najprostszym przypadku (binarnym) wystarczy zapisać je jako 0/1, albo 1/-1.
Dla większej liczby możliwości wprowadza się zmienne pomocnicze - zamiast zapisywać wyjście jedną spośród $K$ wartości, bierzemy $K$ zmiennych binarnych, z których dokładnie jedna jest równa 1 (tzw. one-hot).

# Reprezentacja zmiennych

Zmienne jakościowe mogą być przedstawiane liczbowo. W najprostszym przypadku (binarnym) wystarczy zapisać je jako 0/1, albo 1/-1.
Dla większej liczby możliwości wprowadza się zmienne pomocnicze - zamiast zapisywać wyjście jedną spośród $K$ wartości, bierzemy $K$ zmiennych binarnych, z których dokładnie jedna jest równa 1 (tzw. one-hot).

| Color |
|---|
| Red |
| Red |
| Yellow |
| Green |
| Yellow |



| Red | Yellow | Green |
|---|---|---|
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |

# Notacja

Zmienną wejściową zapisujemy symbolem $X$ (jedno- lub wielowymiarowy wektor). Zaobserwowane wartości zmiennej $X$ będziemy oznaczać $x$.

# Notacja

Zmienną wejściową zapisujemy symbolem $X$ (jedno- lub wielowymiarowy wektor). Zaobserwowane wartości zmiennej $X$ będziemy oznaczać $x$.
Zestawy zmiennych wejściowych zapisujemy macierzowo, np. $N$ wektorów $p$-wymiarowych zapiszemy macierzą wymiaru $N \times p$.

# Notacja

Zmienną wejściową zapisujemy symbolem $X$ (jedno- lub wielowymiarowy wektor). Zaobserwowane wartości zmiennej $X$ będziemy oznaczać $x$.
Zestawy zmiennych wejściowych zapisujemy macierzowo, np. $N$ wektorów $p$-wymiarowych zapiszemy macierzą wymiaru $N \times p$.
Nasz cel - na podstawie realizacji wejściowych zmiennych i przypisanych im wartości zmiennych objaśnianych musimy stworzyć reguły, które pozwolą na przewidywanie wartości dla innych danych wejściowych.

# Pierwsze przykłady - metoda najmniejszych kwadratów

W modelu liniowym zakładamy, że zmienna objaśniana jest zależna od zmiennych wejściowych liniowo, tj. dla wektora $X = (X_1, X_2, \ldots, X_p)^T$ mamy

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j.$$

# Pierwsze przykłady - metoda najmniejszych kwadratów

W modelu liniowym zakładamy, że zmienna objaśniana jest zależna od zmiennych wejściowych liniowo, tj. dla wektora $X = (X_1, X_2, \ldots, X_p)^T$ mamy

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j.$$

Składnik $\hat{\beta}_0$ określa się jako *bias*. Cały model można zapisać w skróconej formie

$$Y = X^T \hat{\beta}.$$

# Metoda najmniejszych kwadratów, cd.

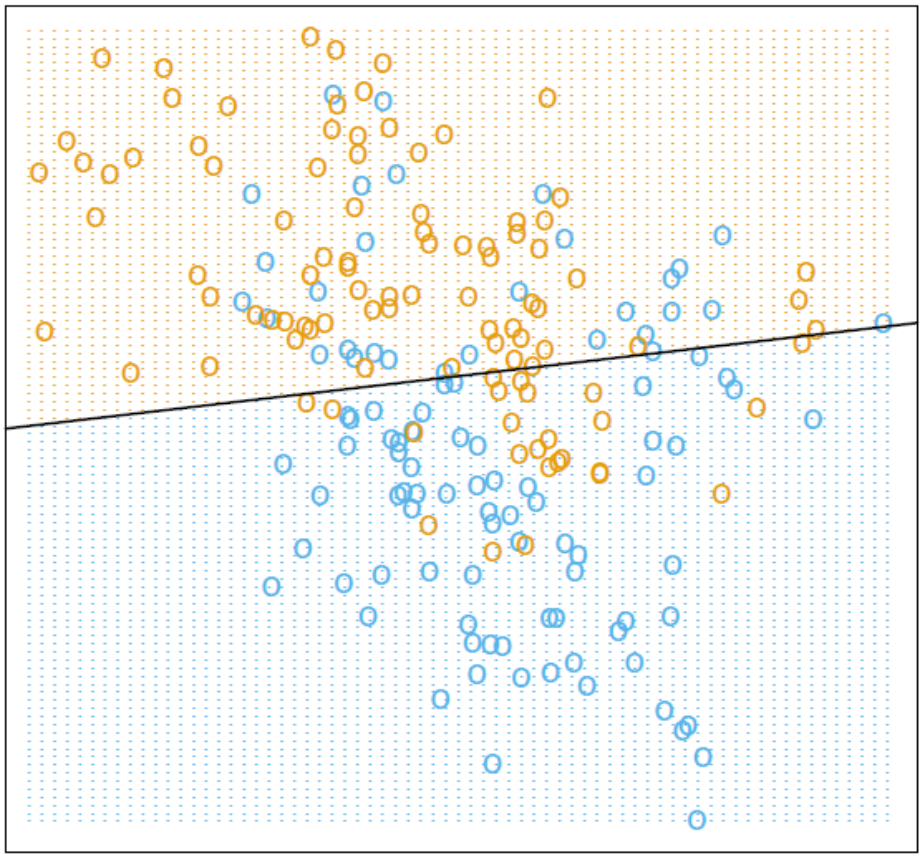
Najpopularniejszą metodą dopasowania współczynników jest minimalizacja sum kwadratów błędów, tj.

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2.$$

Jeżeli macierz $\mathbf{X}^T\mathbf{X}$ jest nieosobliwa, to powyższa wartość jest minimalizowana dla

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

# Metoda najmniejszych kwadratów, ilustracja



Używamy dwóch zmiennych do predykcji - zmienne objaśniane mają wartość 1 dla punktów pomarańczowych, 0 dla niebieskich.

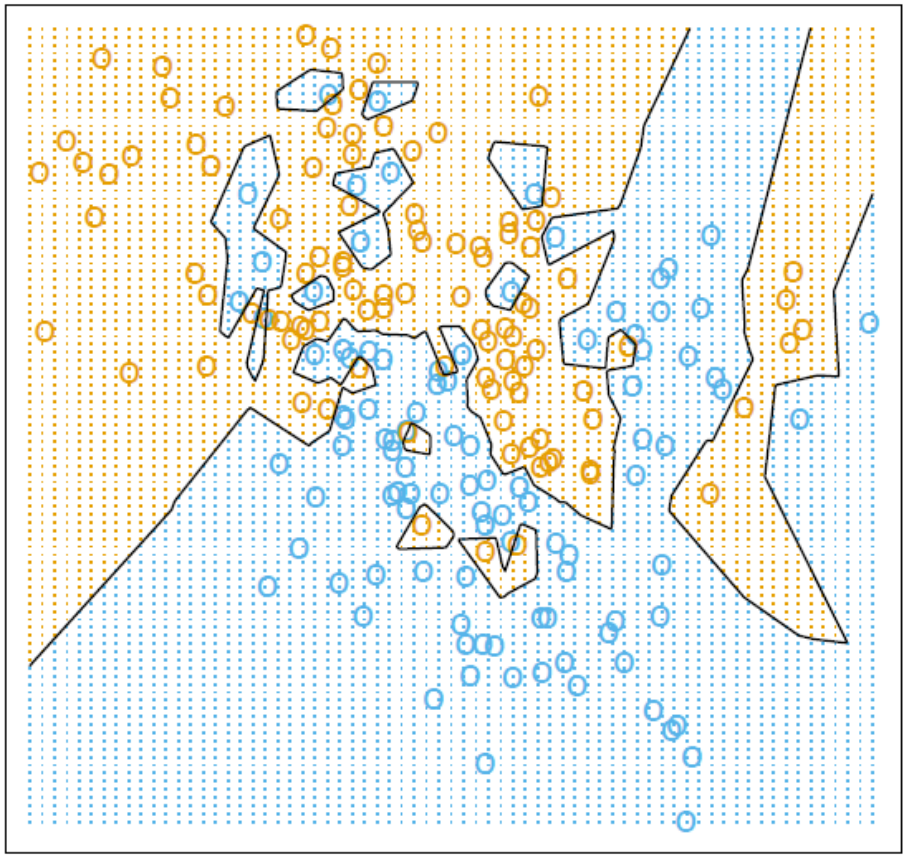
# Pierwsze przykłady - k-Nearest Neighbors

Czasem liniowe (bądź kwadratowe, czy wielomianowe) granice mogą okazać się niewystarczające.

# Pierwsze przykłady - k-Nearest Neighbors

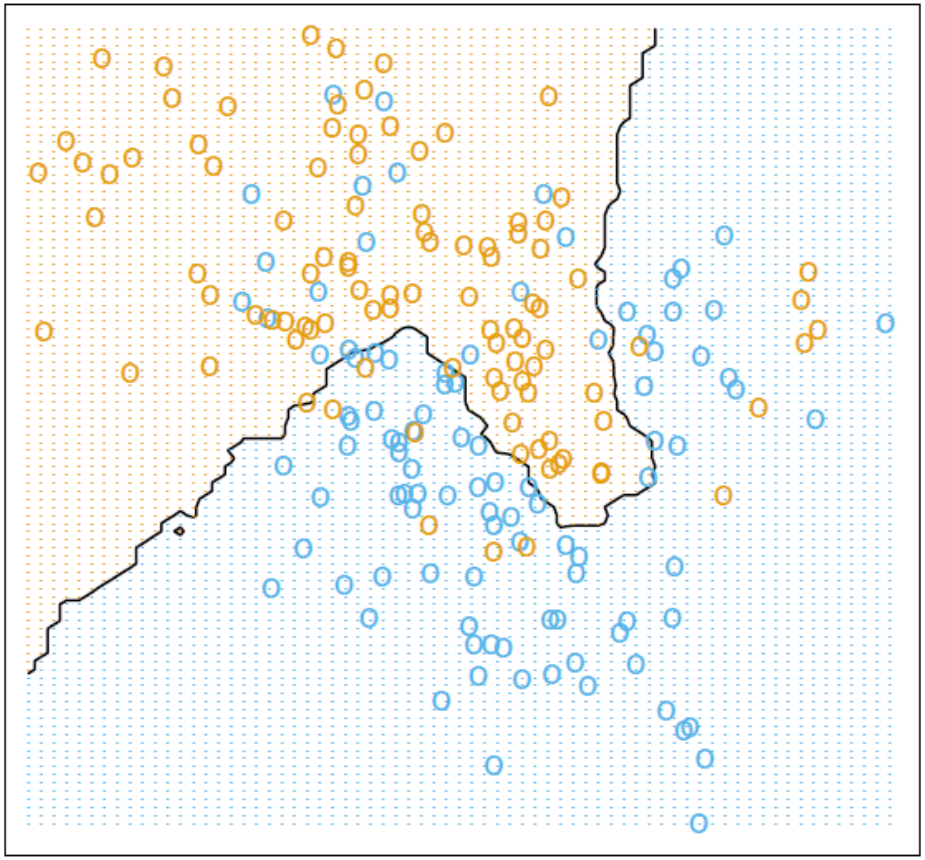
Czasem liniowe (bądź kwadratowe, czy wielomianowe) granice mogą okazać się niewystarczające.
Jednym z nieparametrycznych modeli uczenia z nadzorem jest kNN (k Nearest Neighbours). Model klasyfikuje nowe dane na podstawie pewnej liczby najbliższych sąsiadów wśród zbioru treningowego.

# k-Nearest Neighbors, ilustracja



Używamy tylko najbliższego sąsiada.

# k-Nearest Neighbors, ilustracja



Używamy 15 najbliższych sąsiadów, decyduje większość.

# Teoria decyzji statystycznych

Formalizacja pomysłów na powyższe modele leży w teorii decyzji statystycznych. Dany jest losowy wektor wejściowy $X \in \mathbb{R}^p$ i losowa zmienna wyjściowa $Y \in \mathbb{R}$ o łącznym rozkładzie $P(X, Y)$. Szukamy funkcji $f(X)$, która będzie minimalizowała błąd predykcji dla $Y$.

# Teoria decyzji statystycznych

Formalizacja pomysłów na powyższe modele leży w teorii decyzji statystycznych. Dany jest losowy wektor wejściowy $X \in \mathbb{R}^p$ i losowa zmienna wyjściowa $Y \in \mathbb{R}$ o łącznym rozkładzie $P(X, Y)$. Szukamy funkcji $f(X)$, która będzie minimalizowała błąd predykcji dla $Y$.
Definiujemy funkcję straty $L(Y, f(X))$ która będzie penalizowała błędy - najczęściej kwadratową $L(Y, f(X)) = (Y - f(X))^2$.

# Teoria decyzji statystycznych, cd.

To sugeruje minimalizację $f$ przy pomocy

$$\begin{aligned} EPE(f) &= \mathbb{E}[(Y - f(X))^2] \\ &= \int [y - f(x)]^2 \, P(dx, dy). \end{aligned}$$

# Teoria decyzji statystycznych, cd.

To sugeruje minimalizację $f$ przy pomocy

$$\begin{aligned} EPE(f) &= \mathbb{E}[(Y - f(X))^2] \\ &= \int [y - f(x)]^2 \, P(dx, dy). \end{aligned}$$

Warunkując przez $X$, otrzymujemy

$$EPE(f) = \mathbb{E}_X \mathbb{E}_{Y|X}([Y - f(X)]^2 | X),$$

którego rozwiazaniem jest

$$f(x) = \mathbb{E}(Y|X = x).$$

# Teoria decyzji statystycznych w naszych modelach

kNN implementuje tę metodę, korzystając z uproszczeń:

- Wartość oczekiwana jest przybliżona przez średnią
- Średnia jest liczona tylko na obszarze bliskim danemu $x$.

# Teoria decyzji statystycznych w naszych modelach

Regresja liniowa zakłada, że $f$ ma konkretną postać

$$f(x) = x^T \beta.$$

Po podstawieniu do wcześniejszego wzoru otrzymujemy rozwiązanie na $\beta$:

$$\beta = [\mathbb{E}(XX^T)]^{-1}\mathbb{E}(XY).$$

Tutaj nie warunkujemy po $X$, a jedynie korzystamy z założeń, żeby skorzystać z wartości $X$ w danych treningowych.

# Inne funkcje straty

Co, jeżeli chcielibyśmy zastosować funkcję straty $L_1$, a nie $L_2$?

# Inne funkcje straty

Co, jeżeli chcielibyśmy zastosować funkcję straty $L_1$, a nie $L_2$?
Wtedy szukanym $f$ będzie warunkowa *mediana*.

# Inne funkcje straty

Co, jeżeli chcielibyśmy zastosować funkcję straty $L_1$, a nie $L_2$?
Wtedy szukanym $f$ będzie warunkowa *mediana*.
Co, jeżeli zmienne będą kategoryjne zamiast ciągłych?

# Inne funkcje straty

Co, jeżeli chcielibyśmy zastosować funkcję straty $L_1$, a nie $L_2$?
Wtedy szukanym $f$ będzie warunkowa *mediana*.
Co, jeżeli zmienne będą kategoryjne zamiast ciągłych?
Definiujemy macierz straty (często zero-jedynkową, gdzie każdy błąd karzemy tak samo), i otrzymujemy $f$ postaci

$$f(x) = \max P(g|X = x),$$

tzw. klasyfikator Bayesa.

# Metody dla wyższych wymiarów

*kNN* opiera się na założeniu, że jesteśmy w stanie znaleźć wielu bliskich sąsiadów $X$, by można było wnioskować na ich podstawie. Takie rozumowanie przestaje działać dla większej liczby wymiarów, co nazywamy *przekleństwem wymiarowości*.

# Metody dla wyższych wymiarów

*kNN* opiera się na założeniu, że jesteśmy w stanie znaleźć wielu bliskich sąsiadów $X$, by można było wnioskować na ich podstawie. Takie rozumowanie przestaje działać dla większej liczby wymiarów, co nazywamy *przekleństwem wymiarowości*.
Przykład - dla punktu leżącego w $n$-wymiarowej jednostkowej kostce, jak duże musimy wziąć otoczenie, by objąć 1% lub 10% tej kostki?

# Metody dla wyższych wymiarów

*kNN* opiera się na założeniu, że jesteśmy w stanie znaleźć wielu bliskich sąsiadów $X$, by można było wnioskować na ich podstawie. Takie rozumowanie przestaje działać dla większej liczby wymiarów, co nazywamy *przekleństwem wymiarowości*.
Przykład - dla punktu leżącego w $n$-wymiarowej jednostkowej kostce, jak duże musimy wziąć otoczenie, by objąć 1% lub 10% tej kostki?
Dla 10 wymiarów, odpowiednio 0.63 i 0.80 - nie ma mowy o "lokalności".

# Metody dla wyższych wymiarów

*kNN* opiera się na założeniu, że jesteśmy w stanie znaleźć wielu bliskich sąsiadów $X$, by można było wnioskować na ich podstawie.
Takie rozumowanie przestaje działać dla większej liczby wymiarów, co nazywamy *przekleństwem wymiarowości*.
Przykład - dla punktu leżącego w $n$-wymiarowej jednostkowej kostce, jak duże musimy wziąć otoczenie, by objąć 1% lub 10% tej kostki?
Dla 10 wymiarów, odpowiednio 0.63 i 0.80 - nie ma mowy o "lokalności".
Inny problem - potrzebujemy wykładniczo więcej próbek, by pokryć zbiór treningowy o coraz większej liczbie wymiarów z tą samą gęstością.

# Dobór modelu - przykład

Załóżmy, że model jest deterministyczny i dany wzorem

$$Y = f(X) = e^{-8||X||^2}.$$

Spróbujemy przewidzieć $f(0)$ modelem $1 - NN$ i rozłożymy błąd średniokwadratowy na części:

$$\begin{aligned} MSE(0) &= \mathbb{E}[f(0) - \hat{y}_0]^2 \\ &= \mathbb{E}[\hat{y}_0 - \mathbb{E}(\hat{y}_0)]^2 + [\mathbb{E}(\hat{y}_0) - f(x_0)]^2 \\ &= Var(\hat{y}_0) + Bias^2(\hat{y}_0). \end{aligned}$$

# Dobór modelu, cd.

Dla większej liczby wymiarów bias będzie rósł do 1, natomiast wariancja zacznie spadać (własność tego konkretnego przykładu).

# Dobór modelu, cd.

Dla większej liczby wymiarów bias będzie rósł do 1, natomiast wariancja zacznie spadać (własność tego konkretnego przykładu). Dla modeli liniowych naliczymy dodatkową wariancję (która jednak nie będzie prawie w ogóle rosnąć ze wzrostem wymiaru), ale jeżeli założenia są poprawne, nie będzie biasu (druga część poprzedniego równania zniknie). Oczywiście jeżeli założenia są niepoprawne, to nie jesteśmy w stanie nic powiedzieć o potencjalnej jakości modelu.