

Uczenie z nauczycielem - wprowadzenie, cz.1

Magdalena Buszka

Marzec 2023

Uczenie z nauczycielem:

- Grupa zmiennych z wartościami, które znamy/potrafimy mierzyć (zmiennie niezależne, predyktory)
- Zmienna/e, które są w jakiś sposób związane z predyktorami, część znamy ale w przyszłości nie będziemy mogli/umieli mierzyć ich wartości (zmiennie zależne, zmiennie odpowiedzi)
- Cel - przewidywanie wartości zmiennych zależnych na podstawie zmiennych niezależnych

- Zmienne liczbowe - np. wzrost w cm, ciśnienie atmosferyczne
- Zmienne kategoryczne (dyskretne) - np gatunek irysa, kolor
- Zmienne kategoryczne uporządkowane - np wielkość jako jedna z 3 wartości (mały, średni, duży)

Typ zmiennej zależy wpływa na nazwę zadania predykcji:

- Liczbowa - regresja
- Kategoryczna - klasyfikacja

Zmienne kategoriyczne czasem chcemy kodować przy pomocy liczb.

- Zmienne kategoriyczne binarne - kodowanie przez $\{0,1\}$ lub $\{-1,1\}$
- Zmienne kategoriyczne z większą liczbą stanów - najbardziej popularne kodowanie przez dummy variable. Zmienną z K stanami kodujemy przez K -wymiarowy wektor zero-jedynkowy

- Treningowe
- Walidacyjne (opcjonalnie)
- Testowe

- Regresja liniowa - dużo założeń o strukturze, stabilne ale niekoniecznie bardzo dokładne wyniki
- metoda n najbliższych sąsiadów - bardzo mało założeń o strukturze, niestabilne wyniki, ale często w miarę dokładne

Założmy że mamy jedną obserwację.

Mając wektor zmiennych niezależnych $X^T = (X_1, X_2, \dots, X_p)$ przewidujemy wartość Y przy pomocy modelu

$$\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^p \hat{\beta}_i X_i$$

jeśli do X dopiszemy na początku "zmienną" stale równą 1 to możemy zapisać to macierzowo:

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

- W przestrzeni zmiennych zależnych i niezależnych (X, \hat{Y}) jest pewną hiperpłaszczyzną (jeżeli w X uwzględniliśmy stałą to przechodzi ona przez początkę układu współrzędnych)
- patrząc jako na funkcję - $f(X) = X^T \beta$ jest liniowa, a gradient $f'(X) = \beta$ to wektor wskazujący kierunek największego wzrostu

Dopasowanie modelu liniowego do danych - metoda najmniejszych kwadratów

Minimalizujemy rezydualną sumę kwadratów:

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

macierzowo:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

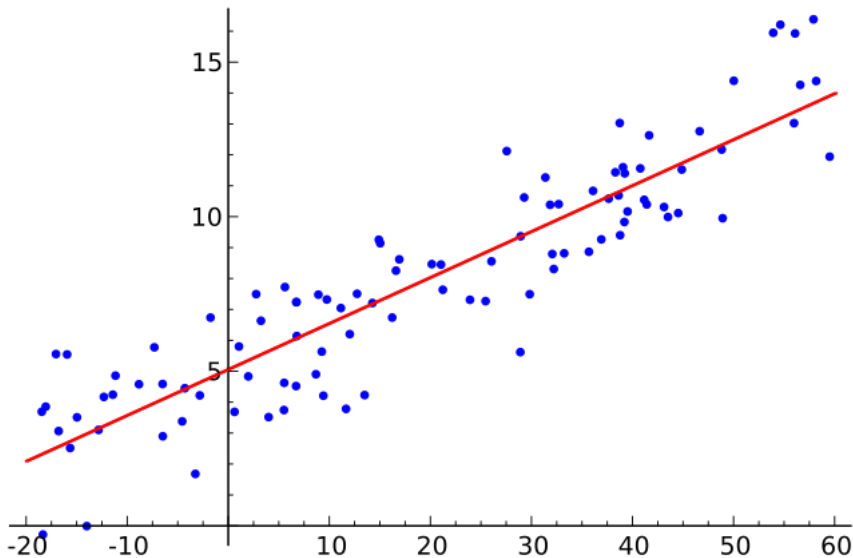
różniczkując ze względu na β otrzymujemy:

$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) = 0$$

gdy $\mathbf{X}^T \mathbf{X}$ jest odwracalne daje to rozwiązanie $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

- liniowość
- niezależność błędów
- stała wariancja błędów
- brak współliniowości kolumn macierzy X

Przykład regresji liniowej



Wartość \hat{Y} wnioskujemy na podstawie wartości zmiennych zależnych obserwacji ze zbioru treningowego \mathcal{T} "najbliższych" obserwacji x .
Dokładniej:

$$\hat{Y} = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$

"bliskość" najczęściej definiujemy w ramach odległości euklidesowej. Dla przypadku binarnej klasyfikacji jest to przypisanie wartości częścię występującej wśród obserwacji w najbliższym otoczeniu

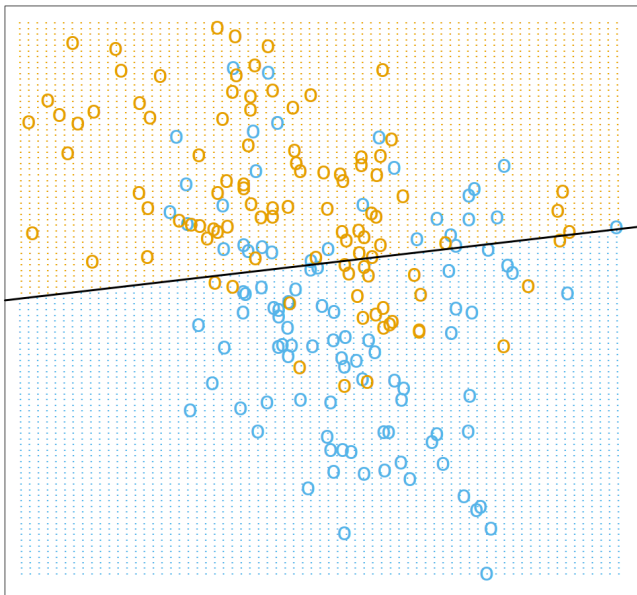
Problem klasyfikacji binarnej o dwóch klasach - "pomarańczowy" i "niebieski".

Zmienne niezależne - pozycja w przestrzeni 2D

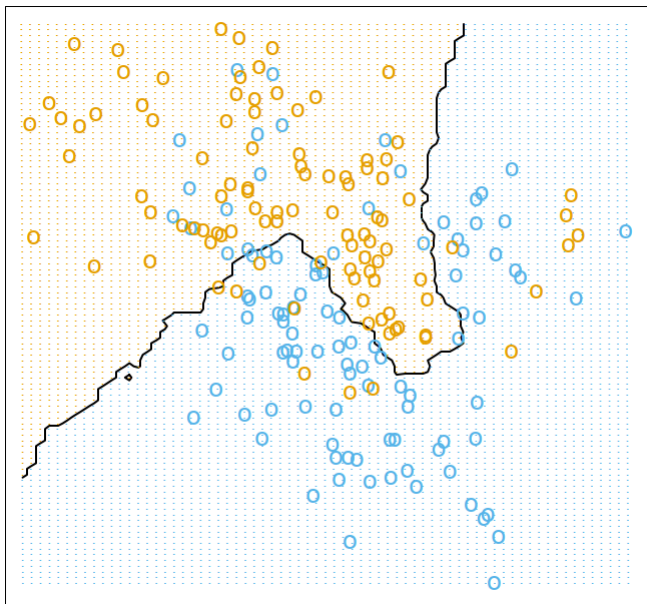
Rozważmy dwie możliwe opcje generowanie zmiennych:

- 1 Wartości wygenerowane są z dwóch wielowymiarowych rozkładów normalnych o różnych średnich i diagonalnych macierzach kowariancji
- 2 Wartości w każdej z klas pochodzą z mieszanki 10 rozkładów normalnych o małych wariancjach i średnich pochodzących z rozkładu normalnego

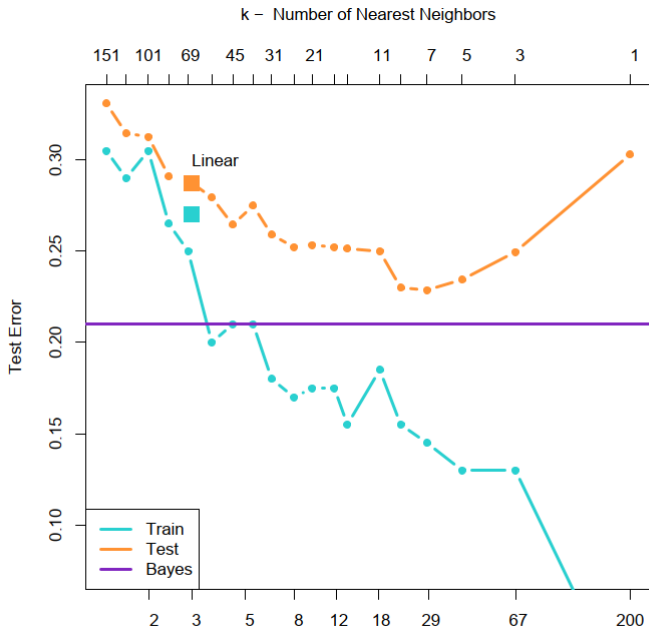
Klasyfikacja na podstawie modelu liniowego



Klasyfikacja na modelu k najbliższych sąsiadów



Podsumowanie



- Model z liniową granicą jest prawie że optymalny dla przypadku 1
- Dla przypadku 2 granica liniowa nie jest dobra, optymalna granica jest nieliniowa i rozłączna. W tym przypadku lepiej radzi sobie model knn.