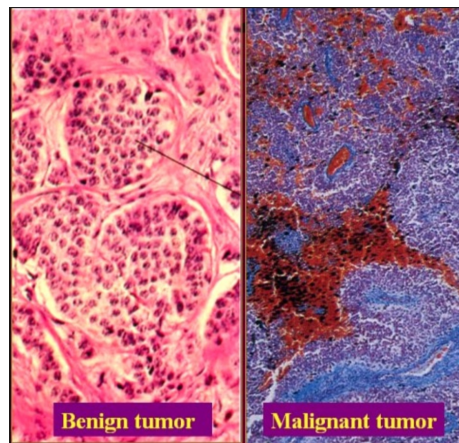


Metoda najmniejszych kwadratów

Zadanie 1.



Celem zadania jest zastosowanie metody najmniejszych kwadratów do predykcji, czy nowotwór jest złośliwy (ang. *malignant*) czy łagodny (ang. *benign*). Nowotwory złośliwe i łagodne mają różne charakterystyki wzrostu. Istotne cechy to m. in. promień i tekstura. Charakterystyki te wyznaczane są poprzez diagnostykę obrazową i biopsje.

Do rozwiązania problemu wykorzystamy bibliotekę `pandas`, typ `DataFrame` oraz dwa zbiory danych:

- `breast-cancer-train.dat`
- `breast-cancer-validate.dat`.

Nazwy kolumn znajdują się w pliku `breast-cancer.labels`. Pierwsza kolumna to identyfikator pacjenta `patient ID`. Dla każdego pacjenta wartość w kolumnie `Malignant/Benign` wskazuje klasę, tj. czy jego nowotwór jest złośliwy czy łagodny. Pozostałe 30 kolumn zawiera cechy, tj. charakterystyki nowotworu.

- Otwórz zbiory `breast-cancer-train.dat` i `breast-cancer-validate.dat` używając funkcji `pd.io.parsers.read_csv` z biblioteki `pandas`.
- Stwórz histogram i wykres wybranej kolumny danych przy pomocy funkcji `hist` oraz `plot`. Pamiętaj o podpisaniu osi i wykresów.

- (c) Stwórz reprezentacje danych zawartych w obu zbiorach dla liniowej i kwadratowej metody najmniejszych kwadratów (łącznie 4 macierze). Dla reprezentacji kwadratowej użyj tylko podzbioru dostępnych danych, tj. danych z kolumn `radius (mean)`, `perimeter (mean)`, `area (mean)`, `symmetry (mean)`.
- (d) Stwórz wektor b dla obu zbiorów (tablicę numpy 1D-array o rozmiarze identycznym jak rozmiar kolumny `Malignant/Benign` odpowiedniego zbioru danych). Elementy wektora b to 1 jeśli nowotwór jest złośliwy, -1 w przeciwnym wypadku.
- (e) Znajdź wagi dla liniowej oraz kwadratowej reprezentacji najmniejszych kwadratów przy pomocy macierzy A zbudowanych na podstawie zbioru `breast-cancer-train.dat`. Potrzebny będzie także wektor b zbudowany na podstawie zbioru `breast-cancer-train.dat`.
Uwaga. Problem najmniejszych kwadratów można rozwiązać stosując równanie normalne lub używając funkcji `scipy.linalg.pinv` lub `scipy.linalg.lstsq`.
- (f) Oblicz współczynnik uwarunkowania, `cond`.
- (g) Sprawdź jak dobrze otrzymane wagi przewidują typ nowotworu (łagodny czy złośliwy). W tym celu pomnóż liniową reprezentację zbioru `breast-cancer-validate.dat` oraz wyliczony wektor wag dla reprezentacji liniowej. Następnie powtórz odpowiednie mnożenie dla reprezentacji kwadratowej. Zarówno dla reprezentacji liniowej jak i kwadratowej otrzymamy wektor p . Zakładamy, że jeśli $p[i] > 0$, to i -ta osoba (prawdopodobnie) ma nowotwór złośliwy. Jeśli $p[i] \leq 0$ to i -ta osoba (prawdopodobnie) ma nowotwór łagodny.

Porównaj wektory p dla reprezentacji liniowej i kwadratowej z wektorem b (użyj reguł $p[i] > 0$ oraz $p[i] \leq 0$).

Oblicz liczbę fałszywie dodatnich (ang. *false-positives*) oraz fałszywie ujemnych (ang. *false-negatives*) przypadków dla obu reprezentacji. Przypadek fałszywie dodatni zachodzi gdy model przewiduje nowotwór złośliwy, gdy w rzeczywistości nowotwór był łagodny. Przypadek fałszywie ujemny zachodzi gdy model przewiduje nowotwór łagodny, gdy w rzeczywistości nowotwór był złośliwy.