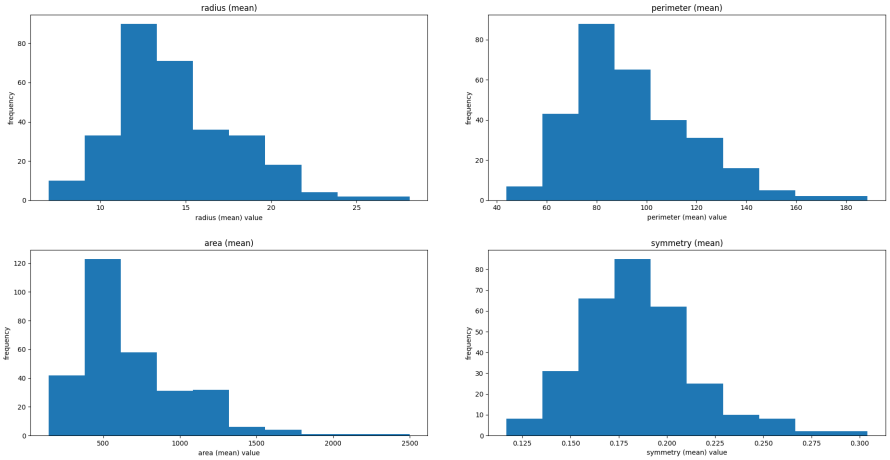
Lab2 - Filip Jedrzejewski

Cel zadania

Celem zadania było zastosowanie metody najmniejszych kwadratów do predykcji, czy nowotwór jest złośliwy czy łagodny. Do rozwiazania problemu wykorzystano biblioteke `pandas`, typ `DataFrame` oraz dwa zbiory danych:
- `breast-cancer-train.dat`
- `breast-cancer-validate.dat`
Sa to zbiory zawierajace wartości dziesieciu cech wykrytych nowotworów oraz to czy dany nowotwór był złośliwy czy łagodny.

Histogramy

Na podstawie powyzszych zbiorów stworzono cztery histogramy roznych kolumn tych danych.



Przygotowanie danych i wyznaczenie wektorów wag

W celu predykcji typu nowotworu stworzono reprezentacje macierzowa obu zbiorów danych dla liniowej i kwadratowej metody najmniejszych kwadratów (łacznie 4 macierze). Do reprezentacji kwadratowej zostały użyte tylko 4 kolumny: `radius (mean)`, `perimeter (mean)`, `area (mean)`, `symmetry (mean)`.
W kolejnym kroku utworzono wektory b dla obu zbiorów, których elementy były równe 1 gdy nowotwór w danym wierszu był złośliwy lub -1 w przeciwnym przypadku.
Nastepnie za pomoca funkcji `scipy.linalg.lstsq` wyznaczono macierz wag dla kwadratowej i liniowej reprezentacji najmniejszych kwadratów.

Współczynniki uwarunkowania

Za pomoca funkcji `numpy.linalg.cond` obliczono współczynniki uwarunkowania zarówno dla liniowej, jak i kwadratowej metody najmniejszych kwadratów:

$$cond(A) = 192499.8$$

$$cond(A_{quad}) = 951673088.7$$

Ocena wyników

Na końcu sprawdzono jakość otrzymanych wyników. W tym celu wymnożono macierze stworzone ze zbioru `breast-cancer-validate.dat` z odpowiadajacymi im macierzami wag. Wynikami tych działań były wektory p i p_{quad}, które zawierały wyniki predykcji. Jeżeli i-ty element wektora p był wiekszy od zera, to i-ta osoba najpewniej miała nowotwór złośliwy w przeciwnym przypadku ($p_i \leq 0$) osoba miała prawdopodobnie nowotwór łagodny.
Porównano otrzymane wyniki z danymi z wektora $b_{validate}$, który przechowywał prawdziwe wyniki. Na tej podstawie obliczono liczbe wyników fałszywie dodatnich, fałszywie ujemnych, prawdziwie dodatnich i prawdziwie ujemnych. Wyniki zapisano w tabeli:

Wyniki dla reprezentacji liniowej

	osoba zdrowa	osoba chora
wynik dodatni	10	58
wynik ujemny	190	2

Wyniki dla reprezentacji kwadratowej

	osoba zdrowa	osoba chora
wynik dodatni	15	55
wynik ujemny	185	5

Wnioski

Na podstawie tych tabel oraz wyznaczonych współczynników uwarunkowania możemy zauważyć, że reprezentacja liniowa metody najmniejszych kwadratów jest w tym zadaniu dokładniejsza i lepsza.