Prep kartkówka ML 1

Czym jest macierz pomyłek (confusion matrix)?

Macierz pomyłek (confusion matrix)

- pozwala obliczyć inne metryki
- **uwaga:** trzeba ogarnąć, gdzie są predykcje, a gdzie wartości prawdziwe

$$P = TP + FN$$

$$N = TN + FP$$

	Actual 1	Actual 0
Predicted 1	True Positive (TP)	False Positive (FP)
Predicted 0	False Negative (FN)	True Negative (TN)

Jest to narzędzie używane do oceny wydajności modelu klasyfikacyjnego. Pokazuje jak dobrze model radzi sobie z różnymi klasami, ułatwia analizę poprawnych jak i błędnych klasyfikacji.
Przedstawiana jest w postaci tabeli gdzie wyróżniamy 4 klasyfikatory: True Positive, False Positive, False Negative, True Negative
Na jej podstawie można obliczyć inne metryki wydajności:

- **Dokładność** (accuracy)

$$(TP + TN)/(TP + TN + FP + FN)$$

- **Precyzja** - mierzy, jak wiele z przewidzianych przypadków pozytywnych (chorych) jest rzeczywiście pozytywnych

$$TP/(TP + FP)$$

- **Czułość** (Recall) - mierzy, jak wiele z rzeczywistych przypadków pozytywnych zostało poprawnie zidentyfikowanych przez model.

$$TP/(TP + FN)$$

- **F1-score**: łączy precyzje i czułość w jedną metrykę

$$2 * (Precision * Recall)/(Precision + Recall)$$

Dlaczego użycie celności (dokładności, accuracy) do oceny klasyfikacji dla zbioru niezbalansowanego jest niepoprawne?

Jeśli użyjemy dokładności do zbioru niezbilansowanego, ponieważ może wprowadzić w błąd, sugerując wysoką wydajność bliską 100% kiedy pomija istotne klasy np. chorych na raka płuc wśród całej populacji.

Przykład niezbalansowanego zbioru: Rozważmy zbiór, w którym 95% osób jest zdrowych, a 5% chorych na raka. Jeśli model klasyfikacyjny przewiduje, że wszyscy są zdrowi, otrzymamy:

- TP=0TP = 0TP=0
- TN=95TN = 95TN=95
- FP=0FP = 0FP=0
- FN=5FN = 5FN=5

$$Dokładność = (TP + TN)/(TP + TN + FP + FN) = 95/100 = 95procent$$

Wydawałoby się że nasz procent jest bardzo dokładny jednak wykrywa chorych.

Jakie najważniejsze hiperparametry mają regresja liniowa i regresja logistyczna?

rodzaj regularyzacji i moc regularyzacji to najważniejsze hiperparametry
Rodzaj regularyzacji

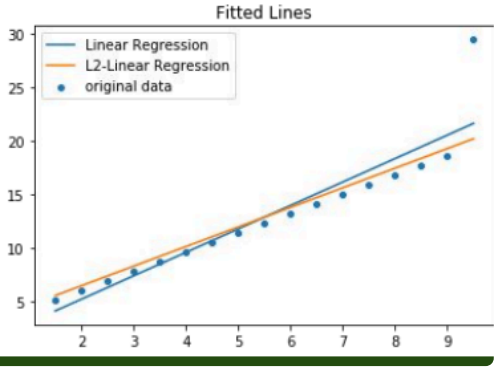
Regresja liniowa

Regularyzacja L2 - współczynnik regularyzacji lambda - zapobiega przeuczeniu dodając karę za duże wartości współczynników

Regularyzacja L2

$$C(X) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{d} w_i^2 = ||y - \hat{y}||_2^2 + \lambda ||w||_2^2$$

- dodajemy do funkcji kosztu sumę kwadratów wag (metryka L2) - formalnie to **ridge regression**
- **penalizuje duże wagi**, które oznaczają zwykle overfitting
- współczynnik regularyzacji λ to **najważniejszy hiperparametr** regresji liniowej



Regularyzacja L1 - hiperparametr premiujący wagi równe dokładnie 0

Regularyzacja L1

$$C(X) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{d} |w_i| = ||y - \hat{y}||_2^2 + \lambda ||w||_1$$

- dodajemy do funkcji kosztu sumę wartości bezwzględnych wag (metryka L1) - formalnie to **LASSO regression**
- penalizuje duże wagi, ale szczególnie **premiuje wagi równe dokładnie 0**
- waga 0 = **selekcja cech (feature selection)**
- **nieróżniczkowalne** - wymaga odpowiedniego solwera
- drugi po L2, albo najważniejszy hiperparametr regresji liniowej

Regresja logistyczna

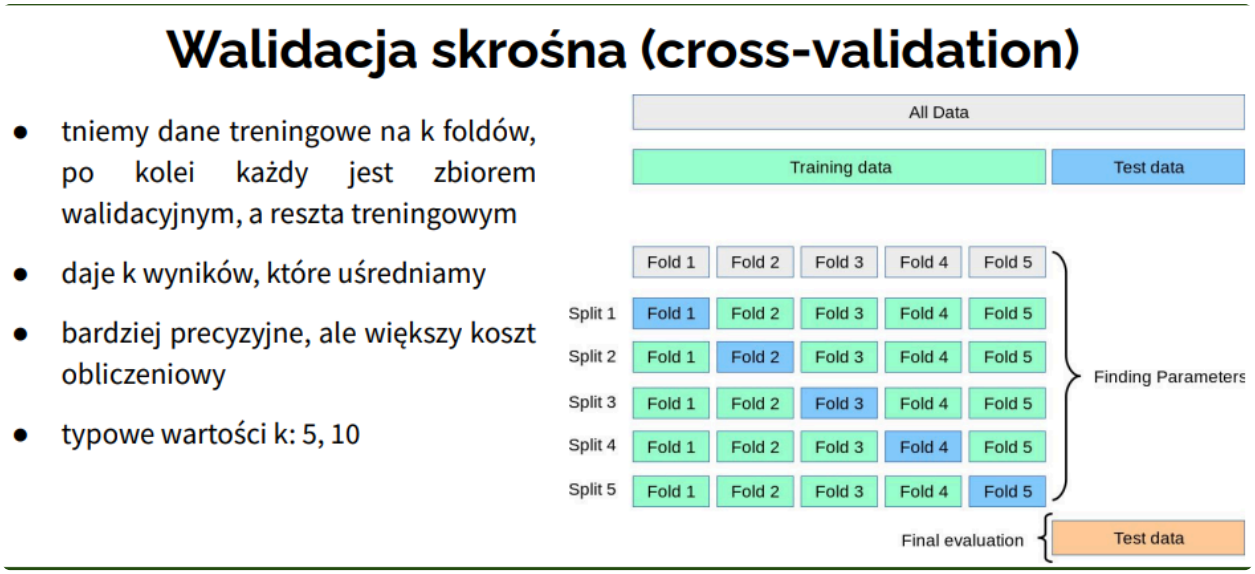
Regularyzacja

- działa analogicznie jak w przypadku regresji liniowej
- dodajemy po prostu dodatkowy czynnik do funkcji kosztu: L1, L2 lub ElasticNet

  $$C(w) = NLL(w) + \lambda ||w||_1 + \gamma ||w||_2^2$$
- L1 i ElasticNet wymagają specjalnych solwerów
- rodzaj regularyzacji i moc regularyzacji to **najważniejsze hiperparametry**

Czym jest walidacja skrośna, czemu jej używamy?

Jest to technika oceny modelu, która polega na podziale zbioru treningowego na k foldów tak że po kolei każdy fold jest zbiorem walidacyjnym, a reszta treningowym. W ten sposób otrzymamy k wyników, które uśrednimy przez co nasz wynik będzie bardziej precyzyjny, a ocena modelu bardziej wiarygodna.
Minusem jest zwiększenie kosztu obliczeniowego k razy.



Czym jest regularyzacja, do czego służy?

Polega na zmniejszeniu pojemności modelu

- zmniejsza **overfitting**
- zmniejsza czułość
  jest to **technika karania dużych wartości parametrów**

Z jakiej funkcji kosztu korzysta regresja liniowa i dlaczego?

MSE Mean Squared Error średni błąd kwadratowy

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

gdzie:

- n to liczba obserwacji,
- y_i to rzeczywista wartość (cel),
- $\hat{y}_i$ to przewidywana wartość przez model.

ponieważ efektywnie mierzy różnice pomiędzy przewidywaną wartością a rzeczywistą.
Zalety: Jest prosta i różniczkowalana

Wady:Kwadratowanie błędów zwiększa wpływ większych błędów na wartość funkcji kosztu, co powoduje że model może się łatwo przeuczyć.

Jakich miar można użyć, do właściwej oceny klasyfikacji niezbalansowanej. Wymień przynajmniej 3 miary.

- precyzja
- czułość (recall)
- f1 score

Adamczyk priv

Jakie znasz klasyfikacje?
co wiesz o niezbalansowanej klasyfikacji?
po co używamy regresji logistycznej? (chodzi o prawdopodobieństwo)
czemu robimy log1p?
co się dzieje jak widzimy taki duży błąd? (tu pokazuje coś) walidacja skrośna
co to jest skrót cv w paramsach?