Wstęp do Uczenia maszynowego

4. czerwca 2024

Powtórka przed egzaminem

Wykład 1 – estymacja parametrów

- Eksperyment losowy
- Próba losowa
- Statystyka
- Estymator
- Estymator nieobciążony
- Estymator efektywny
- Estymator zgodny
- Estymator największej wiarogodności
- Kwantyl
- Estymacja przedziałowa

Wykład 2 – testowanie hipotez

- Hipoteza statystyczna, prosta złożona, parametryczna, nieparametryczna
- Błędy I i II rodzaju
- Moc testu
- Test istotnosci dla wartosci sredniej dla próby normalnej ze znanym σ
- Test istotnosci dla wartosci sredniej dla próby normalnej z nieznanym σ
- Testy istotnosci dla dwóch srednich
- Test Manna-Whitneya/Wilcoxona
- Testy zgodności
- Test zgodności Pearsona
- Test niezależności Pearsona
- Dokładny test Fishera

Wykład 3 – p-wartości

- P-wartość: definicja, relacja z poziomem istotności
- Miara d-Cohena
- Poprawka Bonferoniego
- FDR, procedura Benjaminiego-Hochberga
- Miary korelacji: Pearsona, Spearmana

Wykład 4 – uczenie statystyczne

- Uczenie z nadzorem i bez nadzoru
- Predykcja, wnioskowanie
- Błąd redukowalny i nieredukowalny
- Średni błąd kwadratowy dla regresji
- Kompromis między wariancją a obciążeniem
- Średni błąd kwadratowy dla klasyfikacji
- Czułość, swoistość, precyzja, dokładność
- Klasyfikator bayesowski
- Klasyfikator KNN

Wykład 5 – Regresja liniowa (1)

- Model regresji liniowej
- Założenia modelu regresji liniowej
- Metoda najmnijeszych kwadratów
- Twierdzenie Gaussa-Markowa
- Własności estymatora wektora parametrów β modelu regresji liniowej
- Testowanie istotnosci danego predyktora
- Przedział ufnosci dla estymatora $\hat{\beta}_i$
- Testowanie istotnosci kilku predyktorów
- Overall F test
- odchylenie standardowe składnika resztowego
- Statystyka R^2
- Algorytmy wyboru zmiennych dla regresji linowej

Wykład 6 – regresja liniowa (2)

- Predykcja ze zmiennych jakościowych
- Interakcje zmiennych
- Obserwacje odstające, pojęcie dźwigni
- Współliniowość predyktorów
- Heteroskedastyczność, korelacje reszt

Wykład 7 - Klasyfikacja

- Zagadnienie klasyfikacji
- Regresja logistyczna
- LDA, QDA
- KNN
- Metody oceny jakości klasyfikacji

Wykład 8 – repróbkowanie, wybór modelu

Szacowanie błędu testowego

- Podejście zbioru walidacyjnego
- Walidacja leave one out
- Walidacja krzyżowa

Szacowanie wariancji estymatora parametru

- Bootstrap

Algorytmy selekcji modelu

- Kryteria porównywania modeli o różnej liczbie cech
- Wyczerpujące przeszukiwanie
- Algorytmy zachłanne: przeszukiwanie w przód, wstecz, mieszane

Wykład 9 - Regularyzacja

- Regularyzacja modeli:
  - Regresja grzbietowa
  - Lasso
- Redukcja wymiaru:
  - Analiza skłądowych głównych
  - Regresja składowych głównych

Wykład 10 – metody drzewowe

- Drzewa decyzyjne (zasada działania i metody konstrukcji)
- Bagging
- Lasy losowe
- Boosting

Wykład 11 - SVM

- Klasyfikator o maksymalnym marginesie
- Obserwacje wspierające
- Klasyfikator wektorów wspierających
- Funkcje jądra klasyfikatora
- Maszyny wektorów wspierających
- Analogie do regresji logistycznej

Wykład 12 – Sieci Neuronowe

- Zasada działania perceptronu – problemy z nieliniowymi funkcjami, jak XOR
- Sieci wielowarstwowe, propagacja wsteczna,
- Optymalizacja wag po gradiencie
- Sieci rekurencyjne
- Autoencodery,
- Sieci konwolucyjne (splotowe)

Wykład 13 – PCA i t-SNE

- Analiza składowych głównych jako narzędzie ML bez nadzoru.
- Interpretacja kierunków składowych głównych
- Eksploracja danych przy pomocy PCA
- Metoda t-SNE

Wykład 14 - Klasteryzacja

- Zadanie klasteryzacji
- Podejścia grupowania na k grup na przykładzie k-średnich
- Podejścia bottom-up na przykładzie klastrowania hierarchicznego
- Metody oceny klastrowania:
  - Silhouette score
  - Rand Index

Porzykładowe zadanie z testu

Zadanie 5a Oceń prawdziwość podanych zdań.

T dla metody LOOCV (leave-one-out cross-validation) w zadaniu klasyfikacji: jedyne możliwe wartości estymatora błędu testowego na zbiorach walidacyjnych w poszczególnych iteracjach to 0.0 lub 1.0

F dla metody LOOCV (leave-one-out cross-validation) w zadaniu regresji: jedyne możliwe wartości estymatora błędu testowego na zbiorach walidacyjnych w poszczególnych iteracjach to 0.0 lub 1.0

T jeśli wynikiem k-krotnej walidacji krzyżowej w zadaniu klasyfikacji jest estymator błędu testowego $\hat{b} = 0.47$, a w zbiorach walidacyjnych poszczególnych iteracji jest $n_i = 10$ $(i = 1, \ldots, k)$ obserwacji, to był przynajmniej jeden fold, w którym poprawnie sklasyfikowano nie więcej niż 4 obserwacje

F jeśli wynikiem k-krotnej walidacji krzyżowej w zadaniu regresji jest estymator błędu testowego $\hat{b} = 0.41$, a w zbiorach walidacyjnych poszczególnych iteracji jest $n_i = 10$ $(i = 1, \ldots, k)$ obserwacji, to był przynajmniej jeden fold, w którym uzyskano błąd testowy nie przekraczający 0.4

Zadanie 7a Określ prawdziwość poniższych stwierdzeń. W poniższych odpowiedziach jako przeuczenie (ang. overfitting) rozumiemy sytuację, w której model osiąga wysoką dokładność (ang. accuracy) na zbiorze danych treningowych, ale poziom tej dokładności maleje dla nowych, niebiorących udziału w procesie trenowania modelu, danych testowych.

F Klastrowanie hierarchiczne n-elementowego zbioru obserwacji D wymaga określenia liczby poszukiwanych klastrów (k) przed przystąpieniem do obliczenia macierzy odległości między obserwacjami

F Algorytm k-średnich (ang. k-means) zawsze zbiega do tego samego rozwiązania, niezależnie od tego, jak klastry są inicjowane. Dla ustalenia uwagi rozważamy heurystykę przedstawioną na wykładzie

T Klasyfikator 1-NN (k=1 najbliższych sąsiadów) najczęściej wiąże się z większym ryzykiem przeuczenia (ang. overfitting) niż 10-NN, choć w szczególnych przypadkach może być skuteczniejszy

T Ryzyko przeuczenia (ang. overfitting) w drzewach decyzyjnych wzrasta wraz ze wzrostem głębokości konstruowanego drzewa

Zadanie 12b Rozważmy problem wielokrotnego testowania. Niech $p_1, \ldots, p_m$ bedą p-wartościami kolejnych testów.

_____ Jeśli wylosujemy do odrzucenia hipotezy zerowej testy z z prawdopodobieństwem $\frac{\alpha}{m}$, to będziemy kontrolować FWER na poziomie α.

_____ Jeżli procedura Bonferoniego odrzuca H_0 dla pewnego testu to tak samo będzie w procedurze Benajminiego–Hochberga.

_____ Jeśli wszystkie hipotezy zerowe są prawdziwe to procedury wielokrotnego testowania poprawiają accuracy.

_____ Procedury wielokrotnego testowania pogarszają moc testów.