Elementy teorii uczenia maszynowego
Raport końcowy

Imię Nazwisko

1 Propozycja rozwiązania problemu dziedzinowego metodami uczenia maszynowego

Oddanie tej części raportu będzie wystarczające na ocenę 4.0. Ta część raportu powinna być zwięzła i treściwa. Nie może przekroczyć 4 stron.

1.1 Dziedzina problemu

Krótki opis dziedziny związanej z pracą badawczą lub zawodową studenta.
Maksymalnie 1/2 strony.

1.2 Opis problemu

Krótki opis wyizolowanego problemu z rozważanej dziedziny, który może zostać rozwiązany metodami uczenia maszynowego
Maksymalnie 1/2 strony.

1.3 Dane

Opis danych, które będą używane podczas uczenia, testowania i fazy eksploatacji stworzonego modelu:

1. Cechy (*features*), ich charakter i liczba,
2. Zmienna wyjściowa (*supervision*, *output*, *outcome*), ich charakter (ciągła, dyskretna, jedno- lub wielowymiarowa, itp.) oraz rozmiar (np. liczba etykiet),
3. Sposób pozyskania cech i zmiennych wyjściowych do procesu uczenia i predykcji,
4. Podział danych na zbiór uczący, walidujący i testowy.

Maksymalnie 1 strona.

1.4 Identyfikacja problemu

Formalne zidentyfikowanie problemu:

1. Jaki jest charakter problemu? Regresja, klasyfikacja binarna, wieloklasowa, wieloetykietowa, ranking?
2. Funkcja straty używana do oceny jakości modelu?
3. Jak ma się liczba cech jest do liczby przykładów uczących (znacząco mniejsza/większa, porównywalna)?
4. Czy problem ma charakter inkrementacyjny?

Maksymalnie 1 strona.

1.5 Metoda/Rozwiązanie/Implementacja

Opis możliwego rozwiązania:

1. Jaki algorytm można wykorzystać do rozwiązania problemu?
2. Jeżeli problem jest złożony, czy możliwe jest jego rozbicie na prostsze problemy? Czy taka redukcja pozwoli na optymalne rozwiązanie oryginalnego problemu?
3. Możliwe problemy implementacyjne wynikające z brakujących danych, dużego rozmiaru danych, itp..

Maksymalnie 1 strona.

1.6 Podsumowanie

Na końcu zawsze powinno być jakieś podsumowanie :)
Maksymalnie 1/2 strony.

2 Pytania dodatkowe

Poprawna odpowiedź na każde z poniższych pytań podwyższa ocenę o 1 punkt.

Lista pytań:

1. Rozważmy funkcję straty wartości bezwzględnej (*absolute value loss function*):
$$\ell(y, \widehat{y}) = |y - \widehat{y}|.$$
Pokaż, że jeśli y pochodzi z jakiegoś rozkładu $P(y)$, to optymalną decyzją minimalizującą oczekiwaną wartość funkcji straty jest **mediana**, tzn.:
$$\text{median}(y) = \arg\min_{\widehat{y}} \mathbb{E}_{y\sim P}\left[|y - \widehat{y}|\right].$$

2. Dla klasyfikacji binarnej i zero-jedynkowej funkcji straty, optymalnym klasyfikatorem Bayesowskim jest:

$$h^*(\boldsymbol{x}) = \operatorname{sgn}(\eta(\boldsymbol{x}) - 1/2), \qquad \text{gdzie} \;\; \eta(\boldsymbol{x}) = P(y = 1|\boldsymbol{x}).$$

Wyznacz klasyfikator Bayesowski dla funkcji straty z kosztami klasyfiakcji (*cost-sensitive loss function*):

$$\ell(y, \widehat{y}) = \begin{cases} 0 & \text{jeśli} \;\; y = \widehat{y}, \\ 1 & \text{jeśli} \;\; y = 1, \, \widehat{y} = -1, \\ \beta & \text{jeśli} \;\; y = -1, \, \widehat{y} = 1. \end{cases}$$

Uwaga: dla $\beta = 1$ otrzymujemy błąd zero-jedynkowy, wtedy klasyfikator Bayesowski powinien zgadzać się z poprzednim wynikiem.

3. Naiwny klasyfikator bayesowski (*Naive Bayes*) bazuje na założeniu niezależności cech w danej klasie, tzn. dla dowolnej klasy k i dowolnego $\boldsymbol{x} = (x_1, \ldots, x_m)$,

$$P(\boldsymbol{x}|y = k) = \prod_{j=1}^{m} P(x_j|y = k).$$

Czy to założenie implikuje (lub jest implikowane przez) założenie o ogólnej niezależności cech:

$$P(\boldsymbol{x}) = \prod_{j=1}^{m} P(x_j),$$

bez warunkowania w danej klasie? Odpowiedź uzasadnij (podając kontrprzykład dla odpowiedzi negatywnej, bądź przedstawiając dowód dla odpowiedzi pozytywnej).

4. Czy naiwny klasyfikator bayesowski (*Naive Bayes*) jest klasyfikatorem liniowym, tzn. czy odpowiada funkcji klasyfikującej:

$$h(\boldsymbol{x}) = \operatorname{sgn}(f(\boldsymbol{x})), \qquad \text{gdzie} \;\; f(\boldsymbol{x}) = w_0 + \sum_{j=1}^{m} w_j x_j?$$

Odpowiedź ściśle uzasadnij. Możesz ograniczyć się dla prostoty do cech binarnych, tj. $x \in \{0, 1\}$.

5. Co dzieje się z rozwiązaniem regresji liniowej:

$$\widehat{\boldsymbol{w}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i \right),$$

gdy liczba cech m jest większa od liczby przykładów uczących n? Odpowiedź uzasadnij. Opisz sposób (również z uzasadnieniem), w jaki można temu problemowi zaradzić.

6. Pokaż, że minimalizacja błędu zero-jedynkowe w klasie klasyfikatorów liniowych jest problemem NP-trudnym (wykonaj redukcję do innego problemu NP-trudnego).

7. Pokaż, że omawiane funkcje straty:

   - kwadratowa $\ell(f) = (1-f)^2$,
   - logistyczna $\ell(f) = \log\left(1 + e^{-f}\right)$,
   - zawiasowa $\ell(f) = \max\{0, 1-f\}$,
   - wykładnicza $\ell(f) = e^{-f}$.

   są wypukłe ze wględu na margines f.

8. Pokaż, że jeśli przykłady uczące w klasyfikacji binarnej są generowane z rozkładu $y \sim P(y)$, $y \in \{-1, 1\}$, a następnie $\boldsymbol{x}|y \sim N(\mu_y, \Sigma)$ (tzn. każda z klas ma swoją własną średnią, ale macierz kowariancji jest wspólna), wtedy $\log \frac{\eta(\boldsymbol{x})}{1-\eta(\boldsymbol{x})}$ jest liniową funkcją $\boldsymbol{x}$, gdzie $\eta(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$. Dla prostoty możesz założyć, że Σ jest macierzą jednostkową.

9. Udowodnij, że wszystkie poniższe funkcje straty są skalibrowane ze względu na klasyfikację binarną (*classification calibrated*) i wyprowadź dla każdej funkcji starty postać klasyfikatora bayesowskiego:

   - kwadratowa $\ell(f) = (1-f)^2$,
   - logistyczna $\ell(f) = \log\left(1 + e^{-f}\right)$,
   - zawiasowa $\ell(f) = \max\{0, 1-f\}$,
   - wykładnicza $\ell(f) = e^{-f}$.

10. Udowodnij, że klasa prostokątów na płaszczyźnie ma wymiar Vapnika-Chervonenkisa równy 4.

11. Udowodnij, że klasyfikator bayesowski dla błędu Hamminga i wieloetykietowego błędu zerojedynkowego (ang. *subset 0/1 loss*) wyraża się odpowiednio następującymi wzorami:

    - błąd Hamminga:

      $$\boldsymbol{h}^*(\boldsymbol{x}) = (h_1^*(\boldsymbol{x}), h_2^*(\boldsymbol{x}), \ldots, h_m^*(\boldsymbol{x})),$$

      gdzie $h_i^*(\boldsymbol{x}) = \arg\max_{y_i \in \{0,1\}} \Pr(y_i \,|\, \boldsymbol{x})$, dla $i = 1, \ldots, m$,
    - wieloetykietowy błąd 0/1:

      $$\boldsymbol{h}^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \Pr(\boldsymbol{y} \,|\, \boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \Pr(\boldsymbol{y} \,|\, \boldsymbol{x})$$

      ,

12. Pokaż, że strukturalna metoda wektorów wspierających (ang.*structured support vector machines*) dla funkcji skoringowej $f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} f_i(\boldsymbol{x}, y_i)$ sprowadza się do metody niezależnych klasyfikatorów binarnych (ang. *binary relevance*) ze standardową (tzn. binarną) wersją metody wektorów wspierających. Równoważność należy wykazać na poziomie zastępczych funkcji strat (ang. *surrogate loss functions*).

13. Pokaż, że metoda warunkowych pól losowych (ang.*conditional random fields*) dla funkcji skoringowej $f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} f_i(\boldsymbol{x}, y_i)$ sprowadza się do metody niezależnych klasyfikatorów binarnych (ang. *binary relevance*) ze standardową (tzn. binarną) regresją logistyczną. Równoważność należy wykazać na poziomie zastępczych funkcji strat (ang. *surrogate loss functions*).