Metody systemowe i decyzyjne w informatyce

Laboratorium – Python – Zadanie nr 1

Regresja liniowa

autorzy: A. Gonczarek, J.M. Tomczak, S. Zaręba, M. Zięba, J. Kaczmar

Cel zadania

Celem zadania jest implementacja liniowego zadania najmniejszych kwadratów bez i z regularyzacją ℓ_2 na przykładzie dopasowania wielomianu do danych.

Liniowe zadanie najmniejszych kwadratów

Zakładamy, że dany jest model

$$\overline{y} = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w},$$

gdzie $\mathbf{w} = (w_0\, w_1 \ldots w_{M-1})^{\mathrm{T}}$ jest wektorem parametrów, a $\boldsymbol{\phi}(\mathbf{x}) = (\phi_0(\mathbf{x})\, \phi_1(\mathbf{x}) \ldots \phi_{M-1}(\mathbf{x}))^{\mathrm{T}}$ jest wektorem cech. Na przykład model może być wielomianem M-tego rzędu i wówczas cechy są argumentem podniesionym do kolejnych potęg.

Interesuje nas dopasowanie modelu do dostępnych obserwacji $\mathbf{y} = (y_1\, y_2 \ldots y_N)^{\mathrm{T}}$ oraz $\mathbf{X} = [\mathbf{x}_1\, \mathbf{x}_2\, \ldots \mathbf{x}_N]$. Dalej, przez $\boldsymbol{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1)\, \boldsymbol{\phi}(\mathbf{x}_2) \ldots \boldsymbol{\phi}(\mathbf{x}_N)]^{\mathrm{T}}$ oznaczać będziemy macierz wyliczonych cech dla obserwacji $\mathbf{X}$. Dopasowanie modelu do danych polega na znalezieniu wartości parametrów $\mathbf{w}$. W tym celu będziemy minimalizować funkcję błędu, która określa różnicę między obserwacjami a wartościami zwracanymi przez model. Taką funkcją jest suma kwadratów różnic między predykcjami modelu a obserwacjami, tj.

$$Q(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2. \tag{1}$$

Jest to tzw. **liniowe zadanie najmniejszych kwadratów**.

Zakładając, że rząd $r(\boldsymbol{\Phi}) = M$, policzenie gradientu względem parametrów i przyrównanie go do zera daje jednoznaczne rozwiązanie:

$$\mathbf{w} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y}. \tag{2}$$

Liniowe zadanie najmniejszych kwadratów z regularyzacją ℓ_2

Problemem w liniowym zadaniu najmniejszych kwadratów jest konieczność ustalenia liczby cech, np. stopnia wielomianu. Dobranie zbyt małej lub zbyt dużej liczby skutkować może w otrzymaniu

modelu, który niepoprawnie odzwierciedla charakter szukanej zależności. W tym celu proponuje się ustalenie liczby cech, zazwyczaj dostatecznie dużej, oraz zmodyfikowanie funkcji błędu przez dodanie **regularyzatora** ℓ_2[1]:

$$Q(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \tag{3}$$

gdzie $\lambda > 0$ jest współczynnikiem regularyzacji.

Okazuje się, że zastosowanie regularyzacji nie wymaga założenia o rzędzie macierzy $\mathbf{\Phi}$, tj. dla dowolnego $r(\mathbf{\Phi})$ policzenie gradientu względem parametrów i przyrównanie go do zera daje jednoznaczne rozwiązanie:

$$\mathbf{w} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} + \lambda\mathbf{I}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}, \tag{4}$$

gdzie $\mathbf{I}$ jest macierzą jednostkową.

Selekcja modelu

Dalej będziemy rozpatrywać wielomiany stopnia M. Problem selekcji modelu można rozwiązać na dwa sposoby:

1. Ustalić dostatecznie wysoki stopień wielomianu i zastosować regularyzację ℓ_2.
2. Przyjąć różne modele, tj. różne stopnie wielomianu, a następnie dokonać **selekcji modelu** (ang. *model selection*), tj. wybrać model, dla której wartość funkcji błędu jest najmniejsza.

Do oceny poprawności uzyskanego modelu w procesie selekcji modelu będziemy stosować *błąd średniokwadratowy*:

$$E(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\left(y_n - \overline{y}(\mathbf{x}_n)\right)^2. \tag{5}$$

Wybór stopnia wielomianu

Zakładamy różne wartości $M \in \mathcal{M}$. Przykładowo $\mathcal{M} = \{0, 1, 2, 3, 4, 5, 6, 7\}$, czyli rozpatrywać będziemy wielomianu o stopniu od $M = 0$ do $M = 7$. Uczenie modelu, tj. wyznaczenie parametrów $\mathbf{w}$ wg wzoru (2), odbywa się na podstawie ciągu uczącego (treningowego) $\mathbf{X}$ i $\mathbf{y}$. Natomiast porównanie modeli, tj. różnych stopni wielomianów, odbywa się przy użyciu osobnego zbioru walidacyjnego $\mathbf{X}_{val}$ i $\mathbf{y}_{val}$. Procedura selekcji modelu jest następująca:

1. Dla każdego wielomianu stopnia $M \in \mathcal{M}$ wyznacz wartości parametrów $\mathbf{w}_M$ korzystając z (2) w oparciu o dane $\mathbf{X}$ i $\mathbf{y}$.

[1]Regularyzację ℓ_2 na parametry nazywa się czasem *regularyzacją Tichonowa*.

2. Dla każdego wielomianu stopnia $M \in \mathcal{M}$ o parametrach $\mathbf{w}_M$ wyznacz wartość funkcji błędu E_M o postaci (5) w oparciu o dane $\mathbf{X}_{val}$ i $\mathbf{y}_{val}$.

3. Wybierz ten stopień wielomianu M, dla którego wartość funkcji błędu E_M jest najmniejsza.

Wybór wartości współczynnika regularyzacji

W przypadku stosowania regularyzacji ustalamy dostatecznie duży stopień wielomianu, np. $M = 7$, a następnie wyznaczamy parametry dla różnych wartości współczynnika regularyzacji $\lambda \in \Lambda$. Na przykład $\Lambda = \{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100, 300\}$. Wyznaczenie parametrów $\mathbf{w}$ wg wzoru (4), odbywa się na podstawie ciągu uczącego (treningowego) $\mathbf{X}$ i $\mathbf{y}$. Natomiast porównanie modeli dla różnych wartości λ odbywa się przy użyciu osobnego zbioru walidacyjnego $\mathbf{X}_{val}$ i $\mathbf{y}_{val}$. Procedura selekcji modelu jest następująca:

0. Ustal M.

1. Dla każdej wartości współczynnika regularyzacji $\lambda \in \Lambda$ wyznacz wartości parametrów $\mathbf{w}_\lambda$ korzystając z (4) w oparciu o dane $\mathbf{X}$ i $\mathbf{y}$.

2. Dla każdego wielomianu o parametrach $\mathbf{w}_\lambda$ wyznacz wartość funkcji błędu E_λ o postaci (5) w oparciu o dane $\mathbf{X}_{val}$ i $\mathbf{y}_{val}$.

3. Wybierz te wartości parametrów $\mathbf{w}_\lambda$, dla których wartość funkcji błędu E_λ jest najmniejsza.

Zbiór danych

Dane użyte w zadaniu zostały syntetycznie wygenerowane z następującego obiektu:

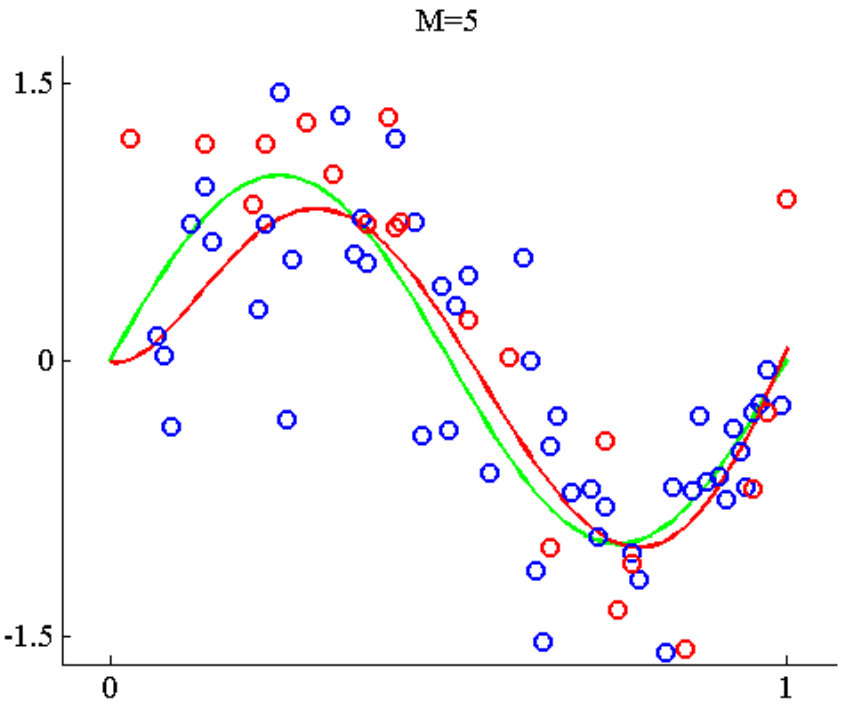
$$y = \sin(2\pi x) + \varepsilon, \tag{6}$$

gdzie $\varepsilon \sim \mathcal{N}(\varepsilon|0, \sigma^2)$ jest szumem gaussowskim, tj. zmienną losową o rozkładzie normalnym i średniej zero. Zbiór danych został podzielony na dwa ciągi treningowe $\mathbf{X}$, $\mathbf{y}$ (odpowiednio po 8 i 50 obserwacji) oraz ciąg walidacyjny $\mathbf{X}_{val}$, $\mathbf{y}_{val}$ (20 obserwacji). Rysunek 1 przedstawia przykładowe dane treningowe (niebieskie punkty) i walidacyjne (czerwone punkty) oraz obiekt (6) (zielona linia) i dopasowany model (czerwona linia).

Testowanie poprawności działania

Do sprawdzania poprawności działania zaproponowanych rozwiązań służy funkcja `main` w pliku `main.py`.

W pliku `main.py` nie wolno czegokolwiek zmieniać ani dopisywać.



Rysunek 1: Zbiór danych oraz przebieg obiektu i modelu.

Instrukcja wykonania zadania

Dodatkowe funkcje, z których należy skorzystać znajdują się w pliku utils.py:

- `polynomial( x, w )` – funkcja zwracająca wartości predykcji `y` dla zadanego `x` oraz wektora wartości parametrów `w` dla domyślnego modelu wielomianu.

Instrukcja:

Należy zaimplementować wszystkie funkcje w pliku `content.py`

1. Zaimplementować funkcję `mean_squared_error` pozwalającą na liczenie średniego błędu kwadratowego (5).
2. Zaimplementować funkcję `design_matrix` liczącą macierz Φ w pliku.
3. Zaimplementować funkcję `least_squares` wyznaczającą rozwiązania liniowego zadania najmniejszych kwadratów.
4. Zaimplementować funkcję `regularized_least_squares` wyznaczającą rozwiązania liniowego zadania najmniejszych kwadratów z regularyzacją ℓ_2.
5. Zaimplementować funkcję `model_selection` dokonującą selekcji modelu dla zadanych wartości $\mathcal{M}$.

6. Zaimplementować funkcję `regularized_model_selection` dokonującą selekcji modelu dla zadanych wartości Λ.

UWAGA! Wszelkie nazwy funkcji i zmiennych w pliku `content.py` muszą pozostać zachowane.

Pytania kontrolne

1. Proszę wyznaczyć rozwiązanie liniowego zadania najmniejszych kwadratów (1).
2. Proszę wyznaczyć rozwiązanie liniowego zadania najmniejszych kwadratów z regularyzacją ℓ_2 (3).
3. Co to jest *overfitting*? Wskazać na przykładzie dopasowania wielomianu.
4. Co to jest *underfitting*? Wskazać na przykładzie dopasowania wielomianu.
5. Co to jest ciąg treningowy, walidacyjny, testowy? Jakie jest ich znaczenie.
6. Co to jest selekcja modelu? W jaki sposób się ją wykonuje? Czy miara oceniająca model może być inna od kryterium uczenia?
7. Które z podejść do selekcji modelu jest prostsze do zastosowania w praktyce i dlaczego?
8. Kiedy liniowe zadanie najmniejszych kwadratów ma jednoznaczne rozwiązanie, a kiedy istnieje wiele rozwiązań? Jak jest w przypadku zadania najmniejszych kwadratów z regularyzacją?
9. Zapisać wektor cech $\boldsymbol{\phi}$ dla wielomianu M-tego rzędu.
10. Co to jest parametr λ? Jak jego wartość wpływa na rozwiązanie?