O dwóch błędach średniokwadratowych

Seweryn Turula

Maj 2022 APF

Statistical learning

Przypuśćmy, że zachodzi relacja

$$Y = f(X) + \varepsilon,$$

gdzie

- Y — zmienna objaśniana (response)
- $X = (X_1, X_2, \ldots X_p)$ — p - zmiennych objaśniających (predictors)
- ε — błąd modelu (error term)
- f — ustalona, ale nieznana nam funkcja X

Po co estymujemy f?

$$\hat{Y} = \hat{f}(X) + \epsilon$$

Są dwa główne powody dla których chcemy estymować funkcję f:

Inference

- wytłumaczenie zależności między X a Y
- które z predyktorów są związane ze zmienną objaśnianą?

Prediciton

- przewidywanie wartości Y
- $\hat{f}$ traktujemy jako *black-box*
- celem jest minimalizacja redukowalnego błędu

Jak estymujemy f?

Podstawą są zaobserwowane przez nas dane (tzw. *training data*), których użyjemy do estymacji f. (inaczej nazywany jako *zbiór uczący*).

Niech x_{ij} oznacza j-ty predyktor dla i-tej obserwacji, gdzie $i \in \{1, 2, \ldots, n\}, j \in \{1, 2, \ldots, p\}$ oraz y_i będzie i-tą obserwacją zmiennej objaśnianej.

Wtedy *training data* jest postaci $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ dla $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.

Metody statystyczne: parametryczne oraz nieparametryczne.

Jak estymujemy f?

Podstawą są zaobserwowane przez nas dane (tzw. *training data*), których użyjemy do estymacji f. (inaczej nazywany jako *zbiór uczący*).

Niech x_{ij} oznacza j-ty predyktor dla i-tej obserwacji, gdzie $i \in \{1, 2, \ldots, n\}, j \in \{1, 2, \ldots, p\}$ oraz y_i będzie i-tą obserwacją zmiennej objaśnianej.

Wtedy *training data* jest postaci $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ dla $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.

Metody statystyczne: parametryczne oraz nieparametryczne.

Jak estymujemy f?

Podstawą są zaobserwowane przez nas dane (tzw. *training data*), których użyjemy do estymacji f. (inaczej nazywany jako *zbiór uczący*).

Niech x_{ij} oznacza j-ty predyktor dla i-tej obserwacji, gdzie $i \in \{1, 2, \ldots, n\}, j \in \{1, 2, \ldots, p\}$ oraz y_i będzie i-tą obserwacją zmiennej objaśnianej.

Wtedy *training data* jest postaci $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ dla $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.

Metody statystyczne: parametryczne oraz nieparametryczne.

Jak estymujemy f?

Podstawą są zaobserwowane przez nas dane (tzw. *training data*), których użyjemy do estymacji f. (inaczej nazywany jako *zbiór uczący*).

Niech x_{ij} oznacza j-ty predyktor dla i-tej obserwacji, gdzie $i \in \{1, 2, \ldots, n\}, j \in \{1, 2, \ldots, p\}$ oraz y_i będzie i-tą obserwacją zmiennej objaśnianej.

Wtedy *training data* jest postaci $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ dla $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.

Metody statystyczne: parametryczne oraz nieparametryczne.

Elastyczność modelu a interpretacja.

Kompromis między elastycznością modelu a interpretowalnością przy użyciu różnych modeli statystycznych. Mówiąc ogólnie, wraz ze wzrostem elastyczności metody zmniejsza się jej interpretowalność.

Miara jakości modelu.

Chcemy określić, jak daleko od oryginalnych zaobserwowanych wartości są wartości otrzymane z modelu statystycznego. Najbardziej znaną miarą jest błąd średniokwadratowy ***(mean squared error)***.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

gdzie $\hat{f}(x_i)$ jest predykcją otrzymaną z $\hat{f}$ dla i-tej obserwacji.

MSE będzie małe, jeśli przewidywane odpowiedzi modelu są bardzo zbliżone do prawdziwych wartości.

Miara jakości modelu.

Chcemy określić, jak daleko od oryginalnych zaobserwowanych wartości są wartości otrzymane z modelu statystycznego. Najbardziej znaną miarą jest błąd średniokwadratowy ***(mean squared error)***.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

gdzie $\hat{f}(x_i)$ jest predykcją otrzymaną z $\hat{f}$ dla i-tej obserwacji.

MSE będzie małe, jeśli przewidywane odpowiedzi modelu są bardzo zbliżone do prawdziwych wartości.

Training MSE i test MSE.

Tak zdefiniowany błąd średniokwadratowy jest obliczany za pomocą ***training data***, które zostały użyte do dopasowania modelu, zatem dokładniej mówiąc powinien być nazywany ***training MSE.***

Ale czy w każdej sytuacji interesuje nas to, jak model sprawdza się i dopasowuje się do wcześniej znanych danych? Może chcielibyśmy wiedzieć, jak nasz model poradzi sobie z wcześniej niewidzianymi danymi?

Training MSE i test MSE.

Tak zdefiniowany błąd średniokwadratowy jest obliczany za pomocą ***training data***, które zostały użyte do dopasowania modelu, zatem dokładniej mówiąc powinien być nazywany ***training MSE.***

Ale czy w każdej sytuacji interesuje nas to, jak model sprawdza się i dopasowuje się do wcześniej znanych danych? Może chcielibyśmy wiedzieć, jak nasz model poradzi sobie z wcześniej niewidzianymi danymi?

Training MSE i test MSE.

Zapis trochę bardziej matematyczny.

Przypuśćmy, że dopasowujemy model statystyczny na naszym wcześniej zebranym *training data* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, oraz otrzymujemy estymator $\hat{f}$ wyjściowej funkcji f.

Następnie, chcemy dowiedzieć się, czy $\hat{f}(x_0)$ jest w przybliżeniu równe y_0, gdzie (x_0, y_0) jest wcześniej niewidzianą *obserwacją testową* nieużytą do dopasowania modelu.

Jeżeli liczba obserwacji (x_0, y_0) jest wystarczająco duża, możemy obliczyć:

$$Ave(y_0 - \hat{f}(x_0))^2,$$

średni kwadrat różnicy predykcji modelu dla obserwacji testowych, nazywany w skórcie ***test MSE*** bądź ***testowy błąd średniokwadratowy.***

Training MSE i test MSE.

Zapis trochę bardziej matematyczny.

Przypuśćmy, że dopasowujemy model statystyczny na naszym wcześniej zebranym *training data* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, oraz otrzymujemy estymator $\hat{f}$ wyjściowej funkcji f.

Następnie, chcemy dowiedzieć się, czy $\hat{f}(x_0)$ jest w przybliżeniu równe y_0, gdzie (x_0, y_0) jest wcześniej niewidzianą *obserwacją testową* nieużytą do dopasowania modelu.

Jeżeli liczba obserwacji (x_0, y_0) jest wystarczająco duża, możemy obliczyć:

$$Ave(y_0 - \hat{f}(x_0))^2,$$

średni kwadrat różnicy predykcji modelu dla obserwacji testowych, nazywany w skórcie ***test MSE*** bądź ***testowy błąd średniokwadratowy.***

Training MSE i test MSE.

Zapis trochę bardziej matematyczny.

Przypuśćmy, że dopasowujemy model statystyczny na naszym wcześniej zebranym *training data* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, oraz otrzymujemy estymator $\hat{f}$ wyjściowej funkcji f.

Następnie, chcemy dowiedzieć się, czy $\hat{f}(x_0)$ jest w przybliżeniu równe y_0, gdzie (x_0, y_0) jest wcześniej niewidzianą *obserwacją testową* nieużytą do dopasowania modelu.

Jeżeli liczba obserwacji (x_0, y_0) jest wystarczająco duża, możemy obliczyć:

$$Ave(y_0 - \hat{f}(x_0))^2,$$

średni kwadrat różnicy predykcji modelu dla obserwacji testowych, nazywany w skórcie ***test MSE*** bądź ***testowy błąd średniokwadratowy.***

Training MSE i test MSE.

Zapis trochę bardziej matematyczny.

Przypuśćmy, że dopasowujemy model statystyczny na naszym wcześniej zebranym *training data* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, oraz otrzymujemy estymator $\hat{f}$ wyjściowej funkcji f.

Następnie, chcemy dowiedzieć się, czy $\hat{f}(x_0)$ jest w przybliżeniu równe y_0, gdzie (x_0, y_0) jest wcześniej niewidzianą *obserwacją testową* nieużytą do dopasowania modelu.

Jeżeli liczba obserwacji (x_0, y_0) jest wystarczająco duża, możemy obliczyć:

$$Ave(y_0 - \hat{f}(x_0))^2,$$

średni kwadrat różnicy predykcji modelu dla obserwacji testowych, nazywany w skórcie ***test MSE*** bądź ***testowy błąd średniokwadratowy.***

Jak obliczyć test MSE?

Możliwe są dwie sytuacje:

- posiadamy dane, których nie użyliśmy do "treningu" modelu
- nie mamy dostępu do takich danych

W pierwszym przypadku rozwiązanie jest łatwe, po prostu wyliczamy *test MSE* za pomocą wcześniejszego wzoru.

W drugim przypadku moglibyśmy wywnioskować, że skoro *test* i *training MSE* są ze sobą powiązane, to powinniśmy wybrać metodę minimalizującą *training MSE.* Ale czy takie podejście jest poprawne?

Jak obliczyć test MSE?

Możliwe są dwie sytuacje:

- posiadamy dane, których nie użyliśmy do "treningu" modelu
- nie mamy dostępu do takich danych

W pierwszym przypadku rozwiązanie jest łatwe, po prostu wyliczamy *test MSE* za pomocą wcześniejszego wzoru.

W drugim przypadku moglibyśmy wywnioskować, że skoro *test* i *training MSE* są ze sobą powiązane, to powinniśmy wybrać metodę minimalizującą *training MSE.* Ale czy takie podejście jest poprawne?

Jak obliczyć test MSE?

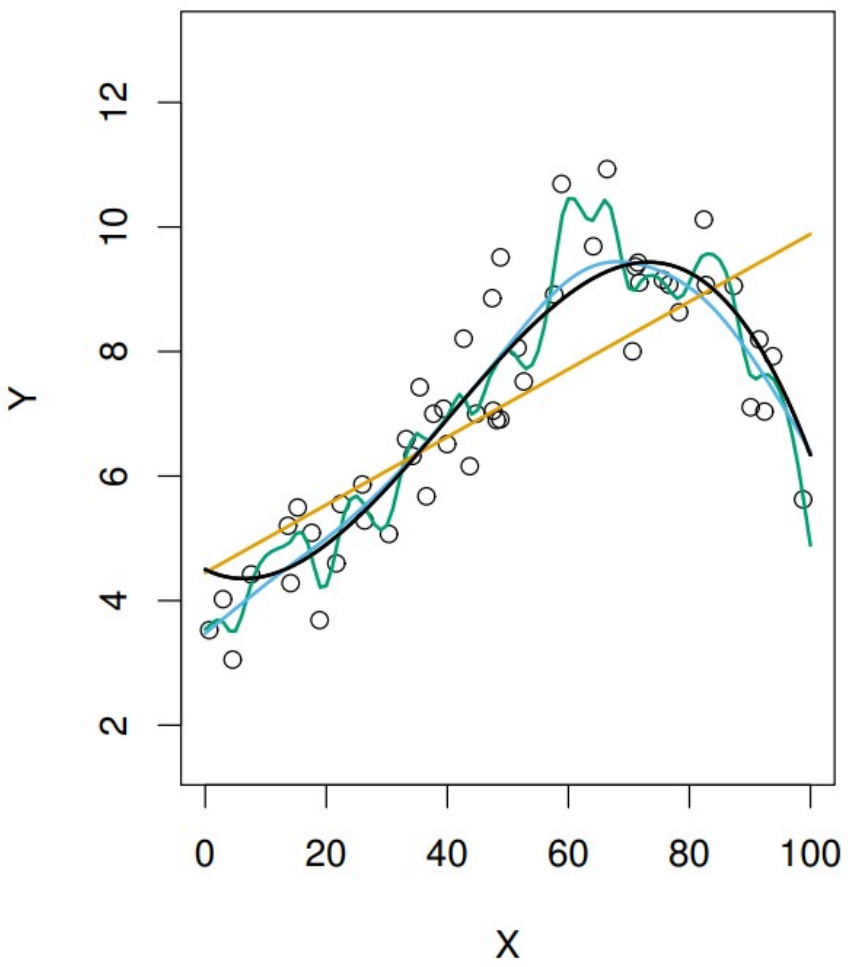
Możliwe są dwie sytuacje:

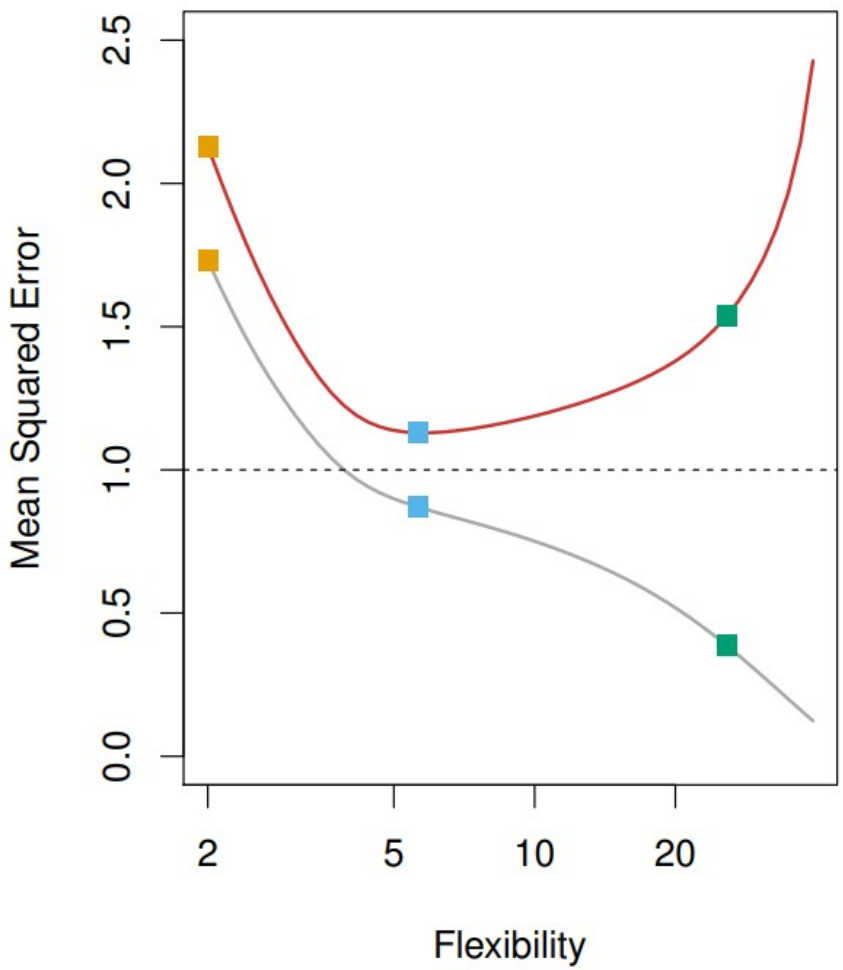
- posiadamy dane, których nie użyliśmy do "treningu" modelu
- nie mamy dostępu do takich danych

W pierwszym przypadku rozwiązanie jest łatwe, po prostu wyliczamy *test MSE* za pomocą wcześniejszego wzoru.

W drugim przypadku moglibyśmy wywnioskować, że skoro *test* i *training MSE* są ze sobą powiązane, to powinniśmy wybrać metodę minimalizującą *training MSE.* Ale czy takie podejście jest poprawne?

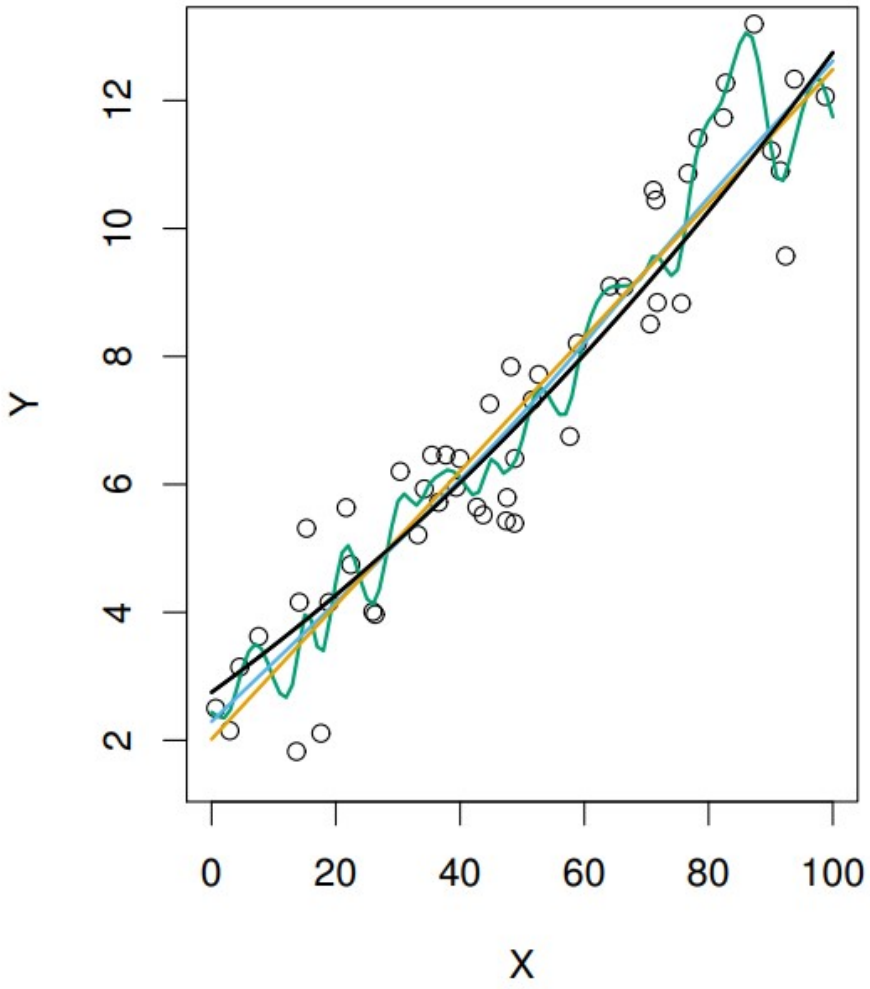
Różnice między test a training MSE (dane nieliniowe)

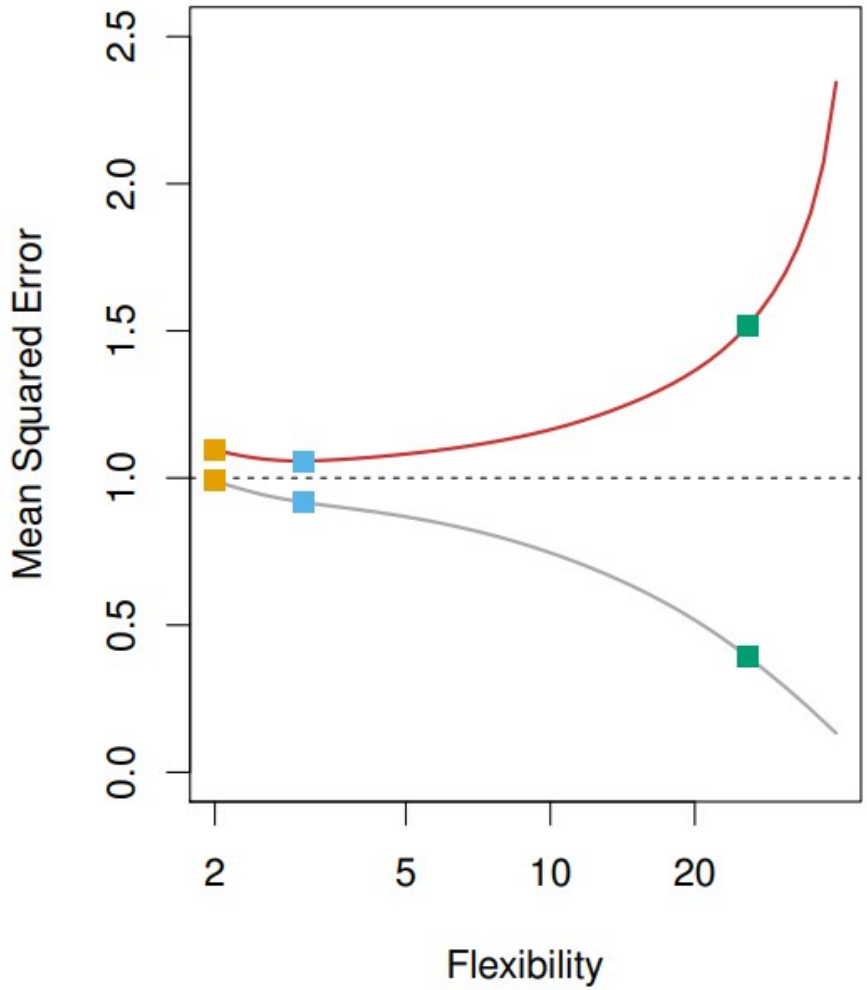




Pomarańczowa linia — regresja liniowa, niebieska i zielona — dwa "smoothing spline". Czerwona krzywa — test, szara — training.

Różnice między test a training MSE (dane liniowe)





Pomarańczowa linia — regresja liniowa, niebieska i zielona — dwa "smoothing spline". Czerwona krzywa — test, szara — training.

Cross-validation

Walidacja krzyżowa lub sprawdzian krzyżowy - metoda statystyczna polegająca na podziale próby statystycznej na podzbiory, a następnie przeprowadzaniu wszelkich analiz na niektórych z nich, tzw. zbiór uczący, podczas gdy pozostałe służą do potwierdzenia wiarygodności jej wyników, tzw. zbiór testowy (branż. zbiór walidacyjny).

- Sprawdzian prosty - najbardziej typowy rodzaj sprawdzianu, w którym próbę dzieli się losowo na rozłączne zbiory: uczący i testowy. Zwykle zbiór testowy stanowi mniej niż 1/3 próby.
- Sprawdzian k-krotny - w tej metodzie oryginalna próba jest dzielona na k podzbiorów.
- Leave-one-out - odmiana sprawdzenia k-krotnego, gdzie elementy podziału są jednoelementowe, tj. N-elementowa próba jest dzielona na N podzbiorów.

The bias-variance trade-off

$$\mathbb{E}(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon),$$

to tak zwane *Expected test MSE w* x_0, które odnosi się do średniego poziomu testowego błędu średniokwadratowego gdybyśmy wielokrotnie wyestymowali wyjściową funkcję f przy użyciu dużych zbiorów uczących.

Dzięki tej zależności, łatwo widać dlaczego wariancja błędu modelu jest ograniczeniem dolnym dla testowego MSE.