Wprowadzenie do uczenia z nadzorem

Mikołaj Słupiński

Czym jest data mining?

Eksploracja danych (ang. data mining) jest dziedziną informatyki, której celem jest dostarczanie algorytmów i technik przetwarzania danych umożliwiających pozyskiwanie wiedzy ze zgromadzonych dużych ilości danych. Metody eksploracji danych opierają się głównie na sztucznej inteligencji i statystyce obliczeniowej.

Uczenie nadzorowane

Uczenie nadzorowane – uczenie maszynowe, które zakłada obecność ludzkiego nadzoru nad tworzeniem funkcji odwzorowującej wejście systemu na jego wyjście.

Rodzaje zmiennych

Ilościowe np. wzrost, waga, temperatura

Rodzaje zmiennych

Ilościowe np. wzrost, waga, temperatura

Jakościowe np. gatunki grzybów, znaki alfabetu

Rodzaje zmiennych

Ilościowe np. wzrost, waga, temperatura

Jakościowe np. gatunki grzybów, znaki alfabetu

Klasyfikacja $f(X) = \hat{G}$

Rodzaje zmiennych

Ilościowe np. wzrost, waga, temperatura

Jakościowe np. gatunki grzybów, znaki alfabetu

Klasyfikacja $f(X) = \hat{G}$

Regresja $f(X) = \hat{Y}$

Modele liniowe

Mając wektor losowy $X^T = (X_1, X_2, ..., X_p)$ chcemy przewidzieć wartości Y, stosując model:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

$$\hat{Y} = X^T \hat{\beta}$$

Jeżeli potraktujemy to jako funkcję $f(X) = X^T \hat{\beta}$ to gradient $f'(X) = \beta$.

Metoda najmniejszych kwadratów

Zdefiniujmy rezidualną sumę kwadratów jako

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2,$$
$$RSS(\beta) = (y - X\beta)^T (y - X\beta),$$

gdzie X jest macierzą wymiaru $N \times p$, natomiast y jest N-wymiarowym wektorem.

Metoda najmniejszych kwadratów

Różniczkując względem β otrzymujemy równania

$$X^T(y - X\beta) = 0.$$

Jeżeli X^TX jest odwracalne, unikalne rozwiązanie zadane jest przez

$$\hat{\beta} = (X^TX)^{-1}X^Ty.$$

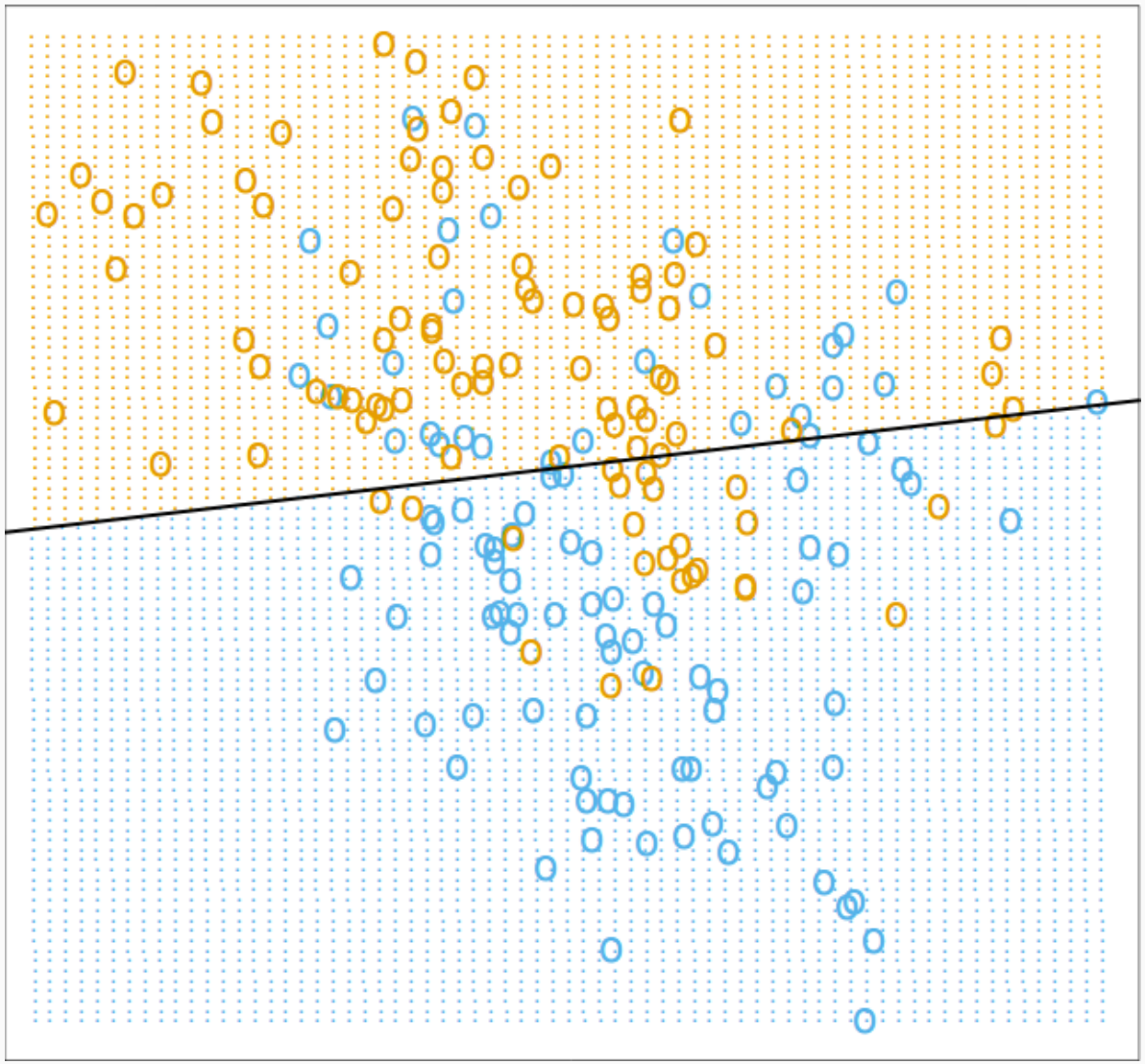
Wówczas wartością predykcji dla x_i jest $\hat{y}_i = \hat{y}(x_i) = x_i^T\hat{\beta}$

Przykład problemu klasyfikacji

Dane składają się z dwóch cech X_1 i X_2. Zmienna wyjściowa G przyjmuje dwie wartości, kolor pomarańczowy i niebieski.

Model regresji liniowej został dopasowany do danych, ze zmienną losową Y przyjmującą wartość 0 dla niebieskich punktów oraz 1 dla pomarańczowych.

Model liniowy w klasyfikacji



Możliwe rozłożenie danych

Scenariusz 1 Dane w każdej klasie zostały wygenerowane przez dwuwymiarowy rozkład Gausa o nieskorelowanych składowych i różnych średnich.

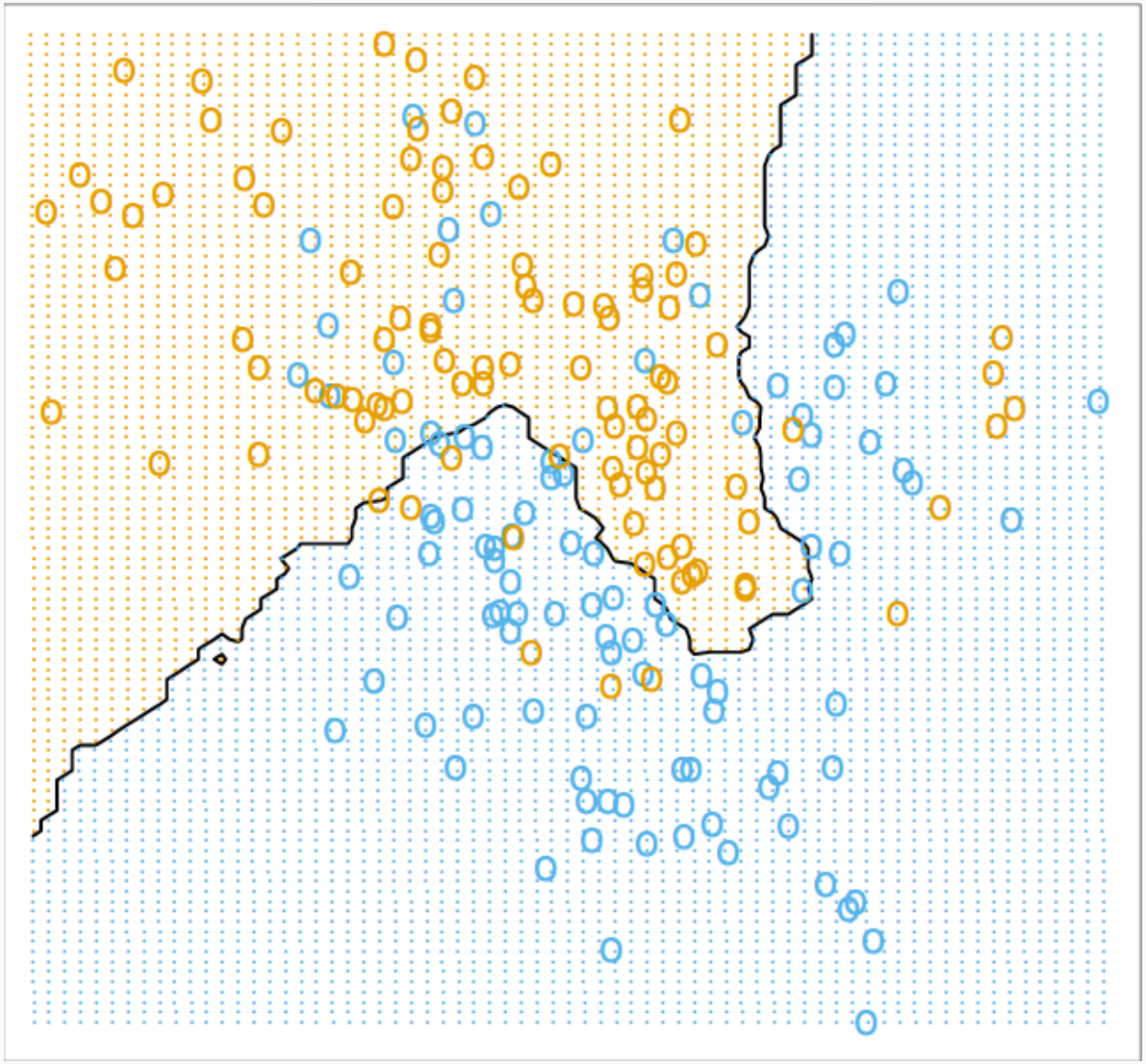
Scenariusz 2 Zbiór treningowy w każdej klasie pochodzi z mieszaniny 10 rozkładów Gaussa o niskiej wariancji i średnich wygenerowanych zgodnie z rozkładem Gaussa.

Klasyfikator KNN

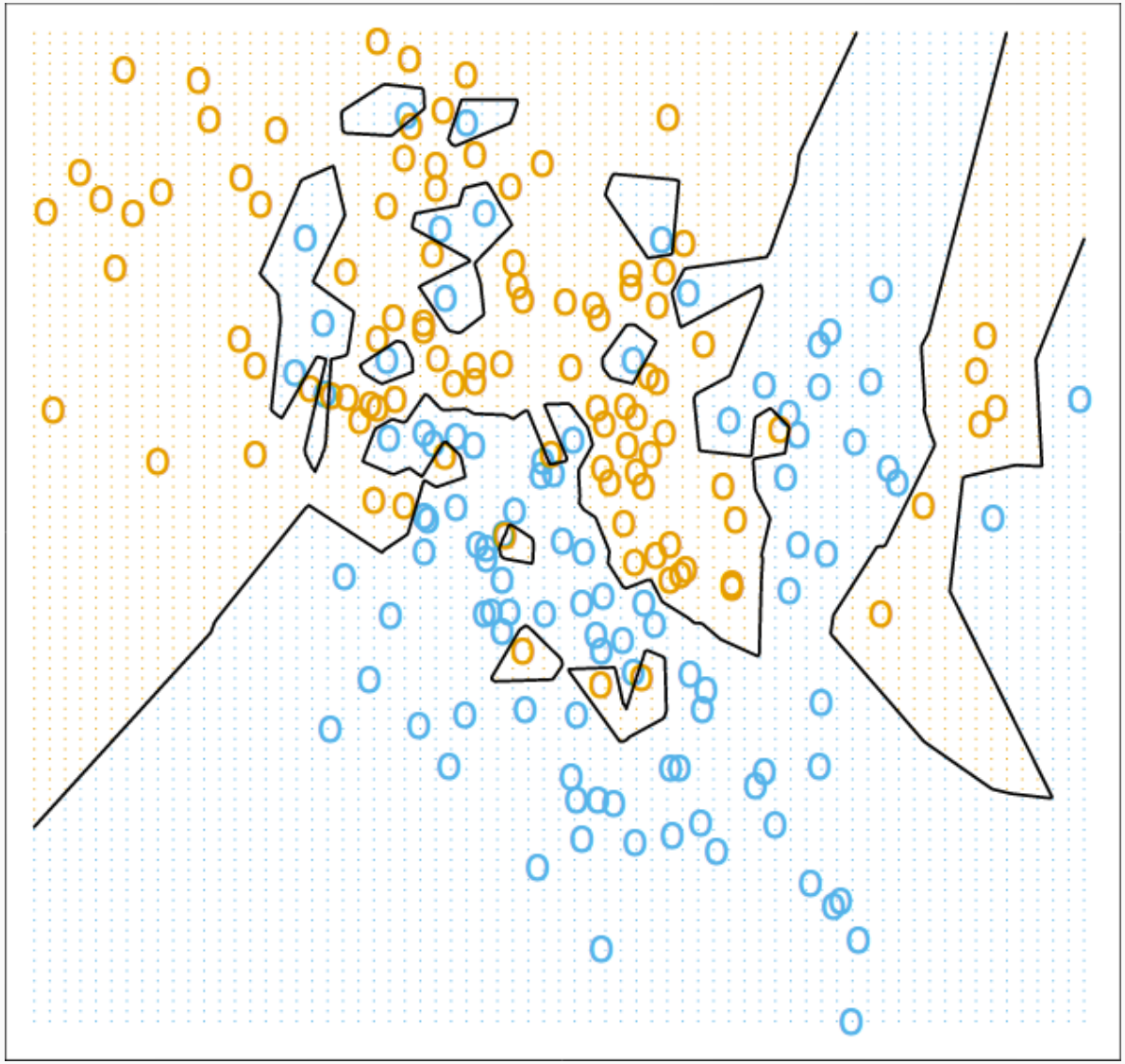
Wynik KNN można zdefiniować jako:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i.$$

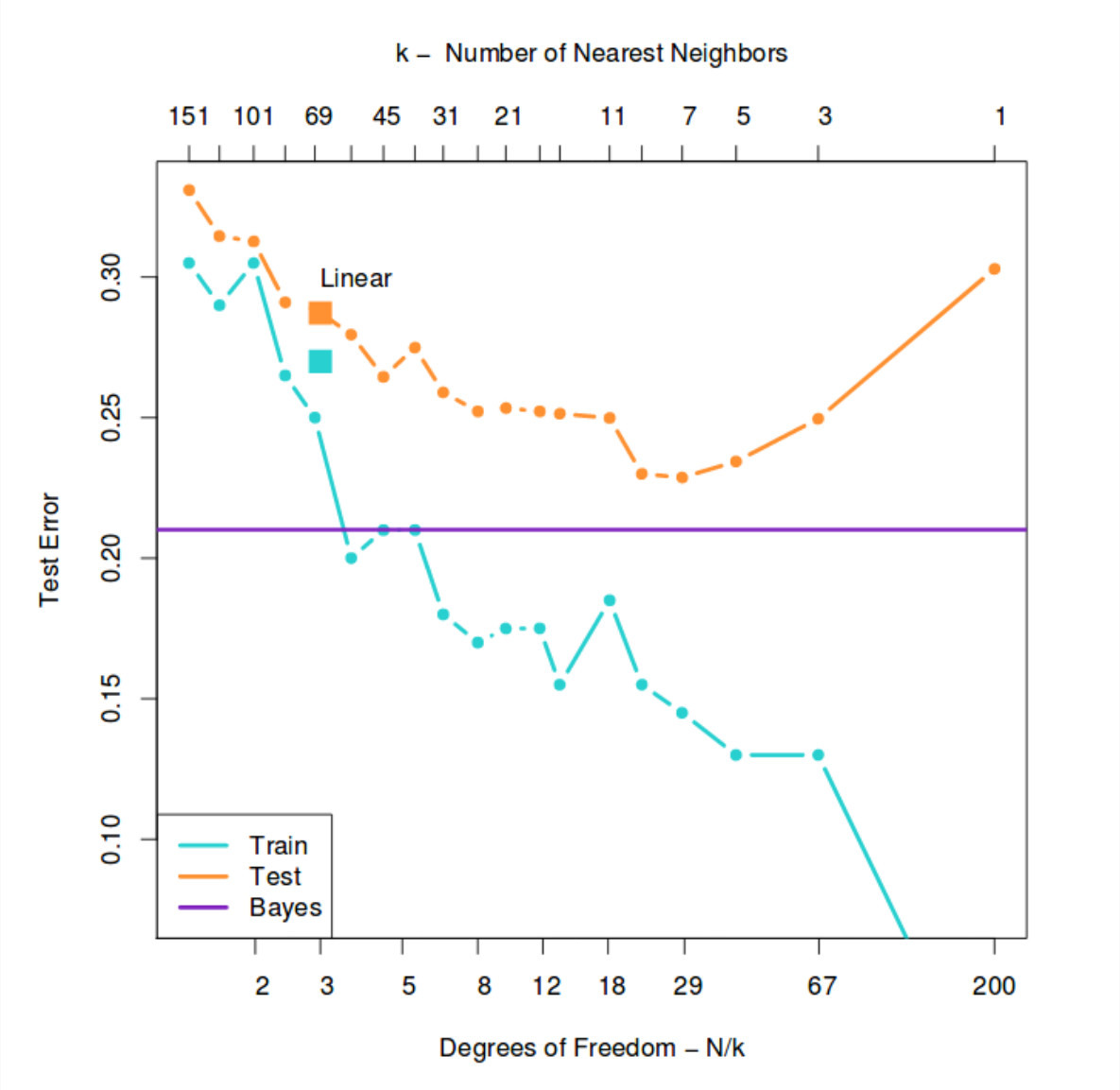
Klasyfikator KNN



Klasyfikator KNN



Klasyfikator KNN



Usprawnienia

- Metody jądrowe używają wag, które zmniejszają się do zera wraz ze wzrostem odległości od punktów,

Usprawnienia

- Metody jądrowe używają wag, które zmniejszają się do zera wraz ze wzrostem odległości od punktów,
- w przestrzeniach o dużej wymiarowości jądra są zmodyfikowane tak, żeby kłaść na pewne zmienne większy nacisk niż na inne,

Usprawnienia

- Metody jądrowe używają wag, które zmniejszają się do zera wraz ze wzrostem odległości od punktów,
- w przestrzeniach o dużej wymiarowości jądra są zmodyfikowane tak, żeby kłaść na pewne zmienne większy nacisk niż na inne,
- lokalna regresja dopasowuje modele liniowe przez lokalnie wyważone najmniejsze kwadraty,

Usprawnienia

- Metody jądrowe używają wag, które zmniejszają się do zera wraz ze wzrostem odległości od punktów,
- w przestrzeniach o dużej wymiarowości jądra są zmodyfikowane tak, żeby kłaść na pewne zmienne większy nacisk niż na inne,
- lokalna regresja dopasowuje modele liniowe przez lokalnie wyważone najmniejsze kwadraty,
- sieci neuronowe składają się z sumy nieliniowo przekształconych modeli liniowych.

Statystyczna teoria podejmowania decyzji

Niech $X \in \mathbb{R}^p$ będzie p-wymiarowym wektorem cech, natomiast $Y \in \mathbb{R}$ zmienną losową oznaczającą wynik z łączonym rozkładem $\mathcal{P}(X, Y)$.

Szukamy funkcji $f(X)$ pozwalającej nam przybliżać wartość Y.

Z tego powodu potrzebujemy *funkcji straty* $L(Y, f(X))$.

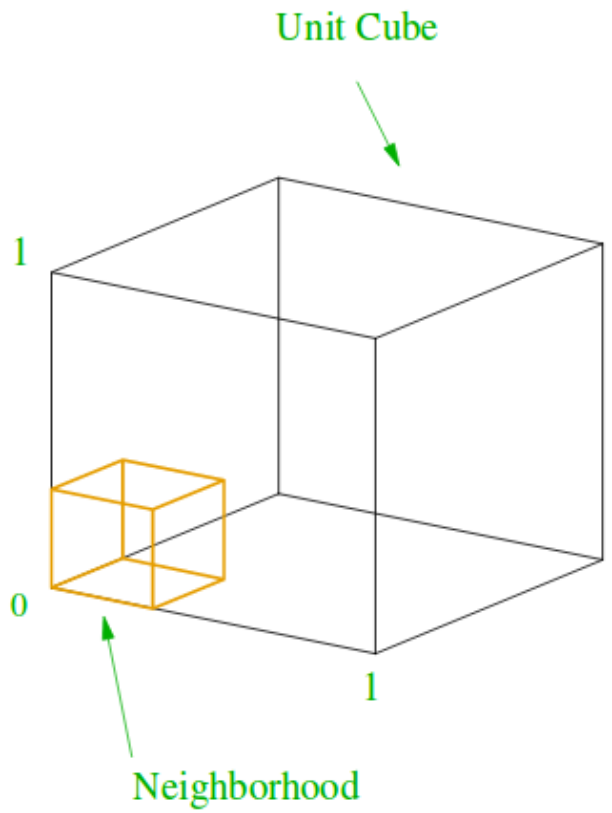
Strata błędu kwadratowego

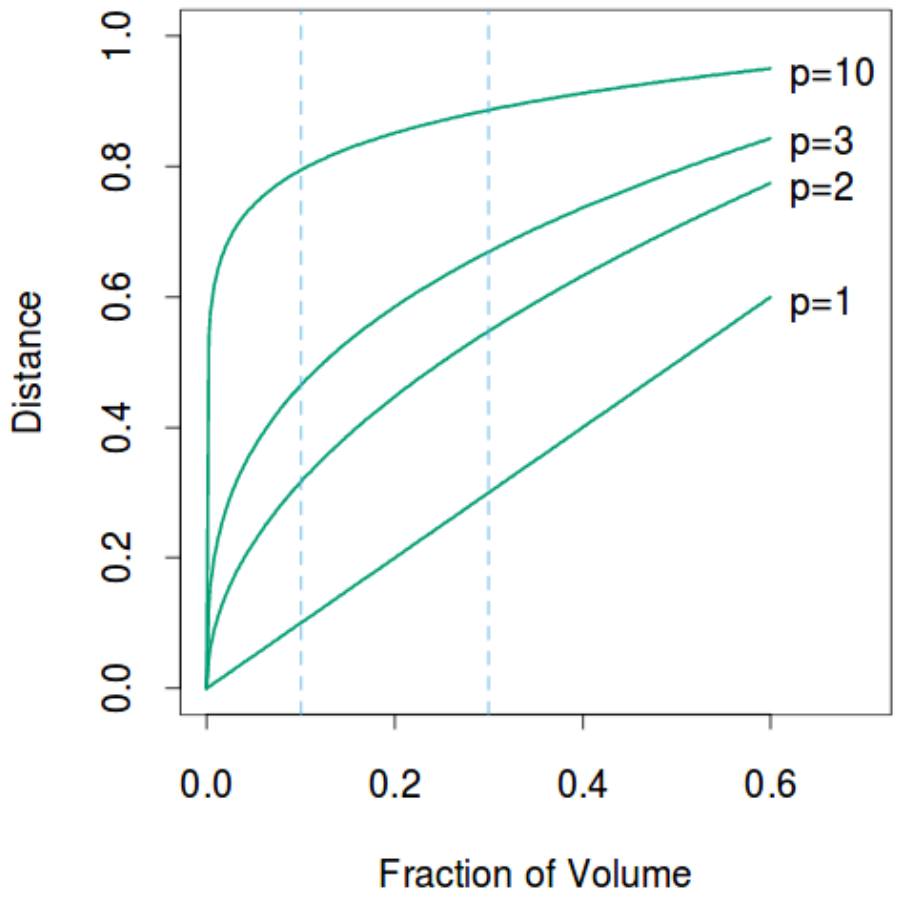
Stratę błędu kwadratowego definiujemy jako

$$L(Y, f(X)) = (Y - f(X))^2$$

$$\begin{aligned} EPE(f) &= \mathbb{E}(Y - f(X))^2 = \int [y - f(x)]^2 \mathcal{P}(dx, dy) \\ &= \mathbb{E}_X \mathbb{E}_{Y|X}([Y - f(X)]^2 | X) \end{aligned}$$

Przekleństwo wymiarowości





Funkcje regresji

Błąd średniokwadratowy prowadzi nas do funkcji regresji

$$f(x) = \mathbb{E}(Y|X = x)$$

.

Przejrzymy teraz inne modele funkcji $f(x)$.

Statystyczny model mieszanego rozkładu $\mathbb{P}(X, Y)$

Przypuśćmy, że dane pochodzą z modelu

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

gdzie błąd losowy ε ma wartość oczekiwaną równą 0 i jest niezależny od X.

Aproksymacja funkcji

Możemy spojrzeć na problem uczenia z nadzorem jako na aproksymację funkcji, gdzie pary (x_i, y_i) należą do $(p+1)$-wymiarowej przestrzeni Euklidesowej.

Przyjęcie takiego podejścia pozwala nam na sparametryzowanie naszych przybliżeń.

Aproksymacja funkcji

Niech θ oznacza zbiór parametrów.

Model liniowy:

$$f(x) = x^T \beta$$
$$\theta = \beta$$

Linear basis expansions:

$$f_\theta(x) = \sum_{k=1}^{K} h_k(x)\theta_k$$

Popularne h_k: $h_k(x) = \frac{1}{1+exp(-x^T \beta_k}$

Estymacja θ

Możemy starać się zminimalizować rezidualną sumę kwadratów:

$$RSS(\theta) = \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2$$

Możemy również zastosować metodę największej wiarogodności:

$$L(\theta) = \prod_{i=1}^{N} f_\theta(y_i)$$

Roughness Penalty

Penalizujemy funkcję RSS(f) :

$$PRSS(f;\lambda) = RSS(f) + \lambda J(f).$$

Np. *cubic smoothing spline*

$$PRSS(f;\lambda) = \sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx.$$

Inne modele

Znamy również inne modele, np.

- metody jądrowe,
- lokalna regresja,
- funkcje bazowe,
- metody słownikowe.

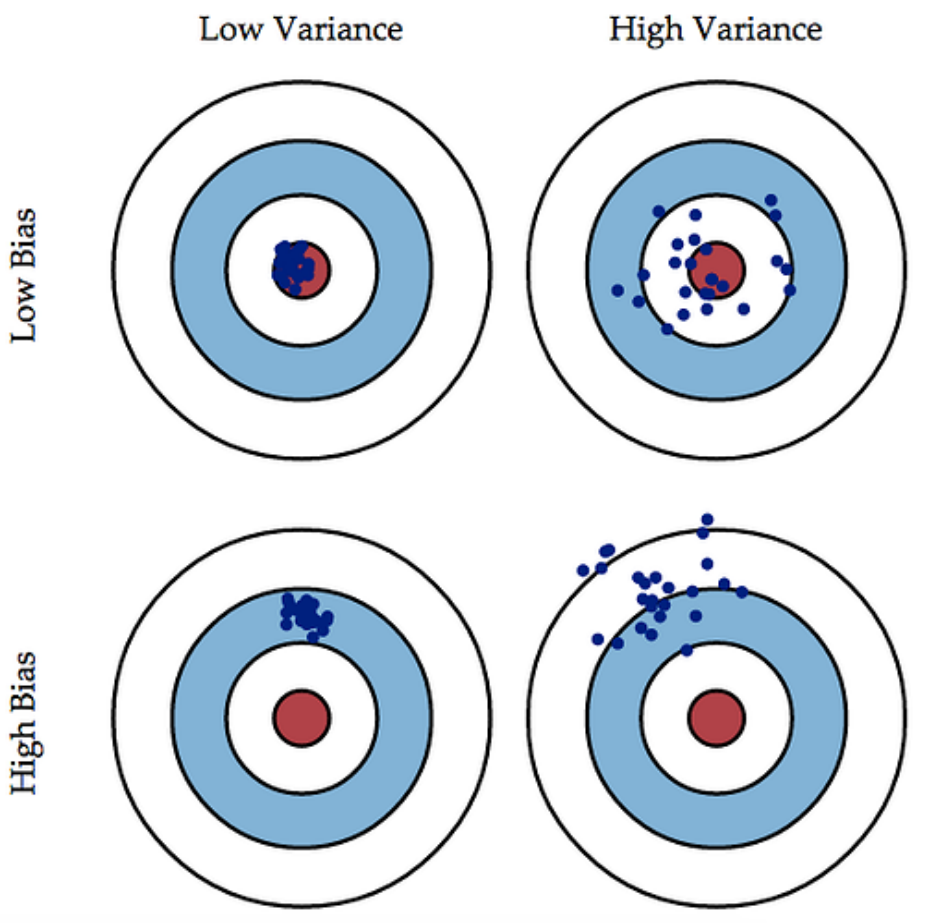
Błąd średniokwadratowy

$$
\begin{aligned}
MSE(x_0) &= \mathbb{E}[y - \hat{y}_0]^2 \\
&= \mathbb{E}[\hat{y}_0 - \mathbb{E}\hat{y}_0]^2 + [\mathbb{E}\hat{y}_0 - y]^2 \\
&= Var(\hat{y}_0) + Bias^2(\hat{y}_0)
\end{aligned}
$$

$$
\begin{aligned}
EPE(x_0) &= \mathbb{E}[(Y - \hat{y}_0)^2 | X = x_0] \\
&= \sigma^2 + Var(\hat{y}_0) + Bias^2(\hat{y}) \\
&= \sigma^2 + MSE(x)
\end{aligned}
$$

gdzie $\hat{y}_0 = \hat{f}(x_0)$, $Y = f(X) + \varepsilon$, $\mathbb{E}\varepsilon = 0$ i $Var(\varepsilon) = \sigma^2$

Wybór modelu i kompromis między wariancją a biasem



Źródło obrazka: https://discuss.analyticsvidhya.com/t/what-is-bias-in-a-machine-learning-algorithm/2171/2