Seminarium magisterskie 2

Teoria decyzji statystycznych, metody używane w wyższych wymiarach

Magdalena Trafidło

3 marca 2023

Rozważamy

$$X \in \mathbb{R}^p, \;\; Y \in \mathbb{R} \sim \mathbb{P}(X, Y).$$

Jak wybrać funkcję $f(X)$ do predykcji Y?
$L(f(X), Y)$ - ustalona funkcja straty, najczęściej

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

Kryterium do wyboru funkcji f:

$$\mathbb{E}PE(f) = \mathbb{E}(Y - f(x))^2 = \int [y - f(x)]^2 \mathbb{P}(dx, dy)$$

Warunkując po X:

$$\mathbb{E}PE(f) = \mathbb{E}_X \mathbb{E}_{Y|X}([Y - f(X)]^2|X).$$

Punktowo:

$$f(x) = \arg\min_c \mathbb{E}_{Y|X}([Y - c]^2|X = x),$$

co daje

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

W metodzie najbliższych sąsiadów

$$\hat{f}(x) = Ave(y_i|x_i \in N_k(x)),$$

gdzie $N_k(x)$ to zbiór k najbliższych sąsiadów punktu x.
Gdy $N, k \to \infty, k/N \to 0$

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

Regresja liniowa:

$$f(x) \approx x^T \beta,$$

teoretyczne rozwiązanie:

$$\beta = [E(XX^T)]^{-1} E(XY).$$

$\hat{\beta} = (XX^T)^{-1} XY$ – zastąpienie wartości oczekiwanej średnia.

k najbliższych sąsiadów (kNN) a metoda najmnmiejszych kwadratów (LS):
- LS i kNN: warunkowe wartości oczekiwane przybliżane średnimi,
- LS: $f(x)$ szacowane globalnie,
- kNN: $f(x)$ szacowane funkcją lokalnie stałą.

Co jeśli zastąpimy L_2 przez L_1?

$$L_1 : E|Y - f(x)|,$$

$$\hat{f}(x) = \text{median}(Y|X = x).$$

Co jeśli chcemy wyznaczyć predykcję zmiennej kategorycznej G?
Funkcja straty:

$$\text{macierz } L \text{ rozmiaru } K \times K,\ K = |G|,$$

$$L(k,k) = 0,$$

$$\text{zazwyczaj } L(k,l) = 1.$$

$$\mathbb{E}(PE) = \mathbb{E}[L(G, \hat{G}(X))] = \mathbb{E}_X \sum_{k=1}^{K} L(G_k, \hat{G}(X))\mathbb{P}(G_k|X).$$

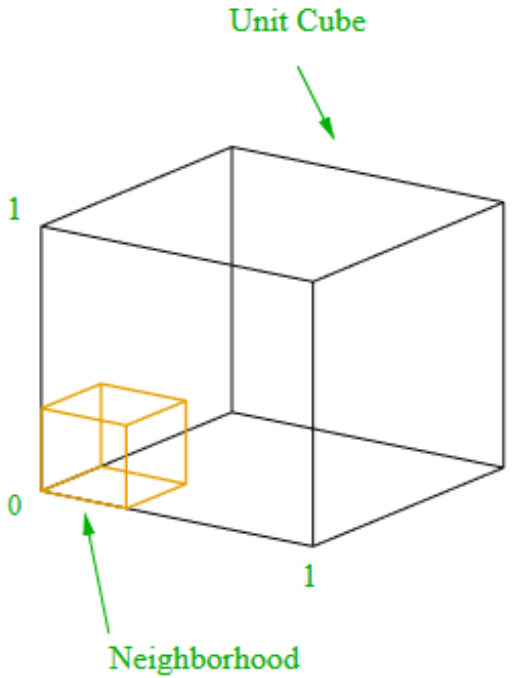
Minimializując punktowo:

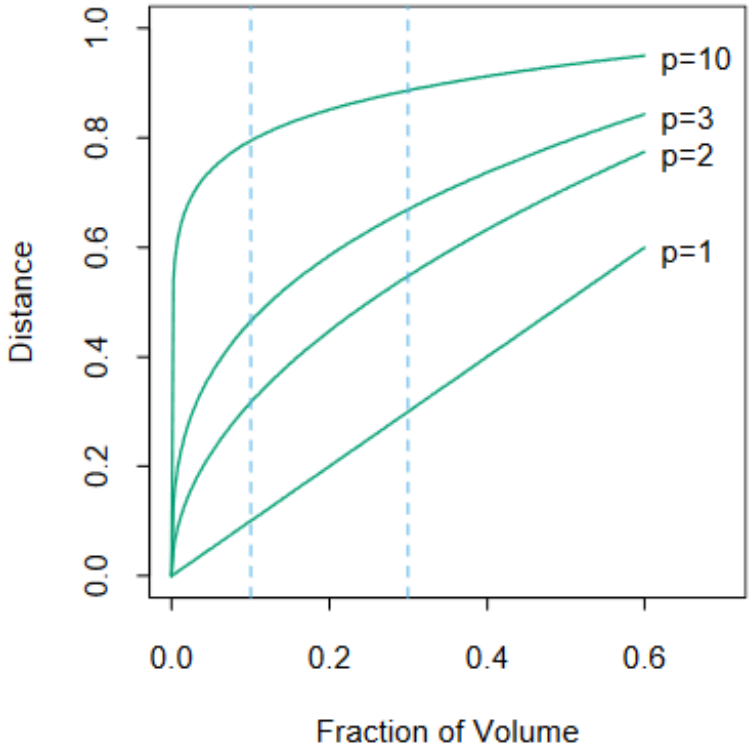
$$\hat{G}(X) = \arg\min_{g \in G} \sum_{k=1}^{K} L(G_k, \hat{G}(X)) \mathbb{P}(G_k | X = x).$$

Gdy L macierz 0-1:

$$\hat{G}(X) = \arg\min_{g \in G} [1 - \mathbb{P}(G_k | X = x)].$$

Rozważmy jednostkowy hipersześcian w $\mathbb{R}^p$ i punkty równomiernie w nim rozmieszczone.
Niech r - ułamek obserwacji, który chcemy uchwycić w sąsiedztwie.
Wtedy $e_p(r) = r^{1/p}$ - oczekiwana długość boku hipersześcianu.
Jeśli $p = 10$ to $e_{10}(0.01) = 0.63$, $e_{10}(0.1) = 0.80$.


Unit Cube
1
0
1
Neighborhood


Distance
0.0
0.2
0.4
0.6
0.8
1.0
p=10
p=3
p=2
p=1
0.0
0.2
0.4
0.6
Fraction of Volume

Rozważmy p-wymiarową hiperkulę ze środkiem w 0.

$$d(p, N) = \left(1 - \left(\frac{1}{2}\right)^{1/N}\right)^{1/p}.$$

$$N = 500, p = 10, d(p, N) \approx 0.52.$$

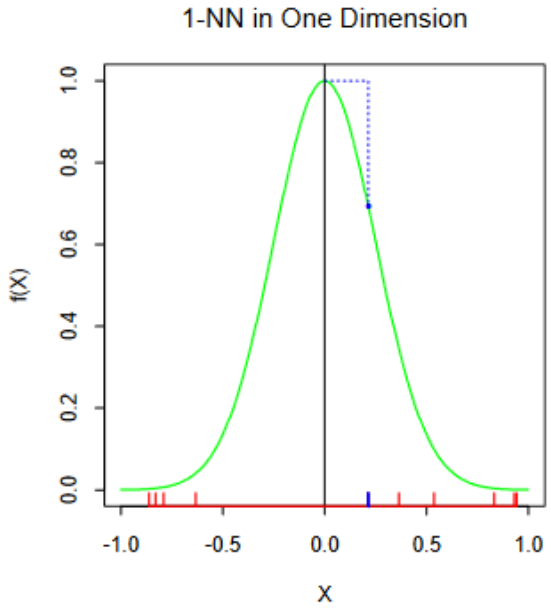
Przykład:

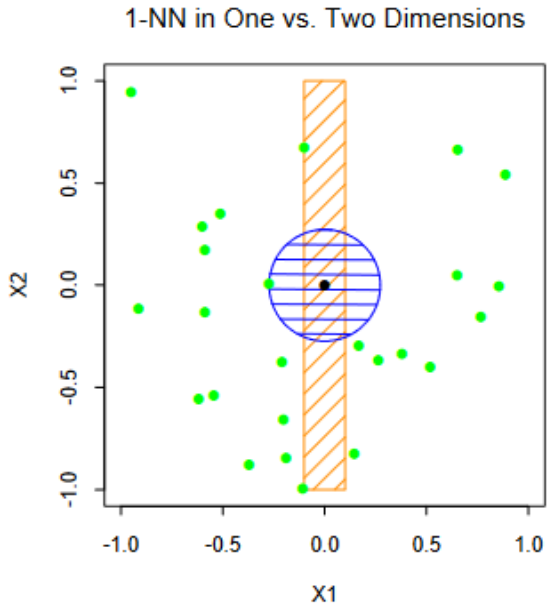
$$N = 1000, x_i \sim U[-1, 1]^p,$$

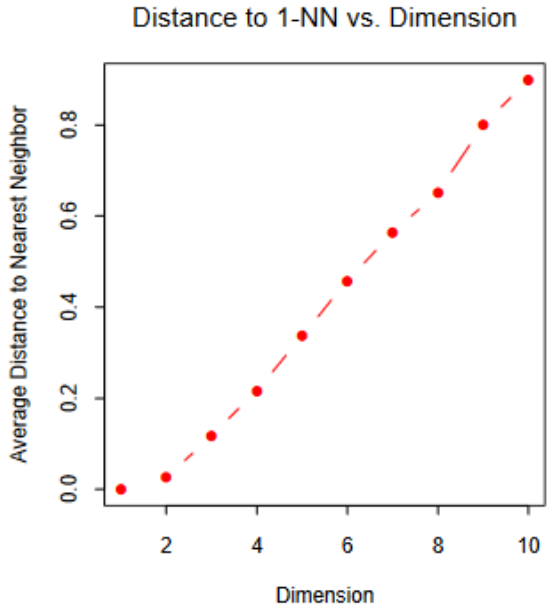
$$Y = f(X) = e^{-8||X||^2}.$$

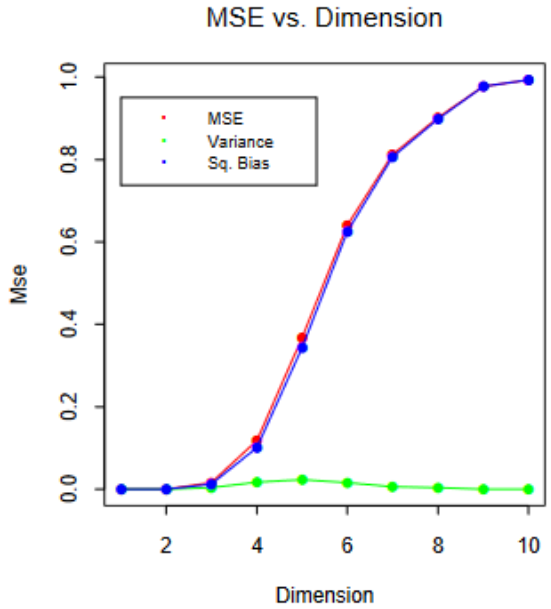
1NN dla $x_0 = 0$:

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


1-NN in One Dimension
f(X)
X
0.0
0.2
0.4
0.6
0.8
1.0
-1.0
-0.5
0.0
0.5
1.0


1-NN in One vs. Two Dimensions
X2
X1
1.0
0.5
0.0
-0.5
-1.0


Distance to 1-NN vs. Dimension
Average Distance to Nearest Neighbor
Dimension
0.0
0.2
0.4
0.6
0.8
2
4
6
8
10


MSE vs. Dimension
Mse
Dimension
MSE
Variance
Sq. Bias
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10

$$Y = X^T\beta + \epsilon,$$

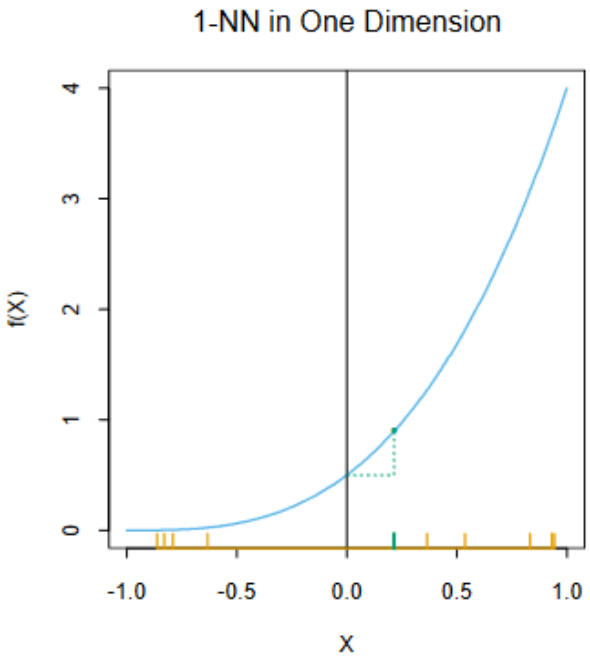
$$\hat{y}_0 = x_0^T\hat{\beta} = x_0^T\beta + \sum_{i=1}^{N} \ell_i(x_0)\epsilon_i,$$

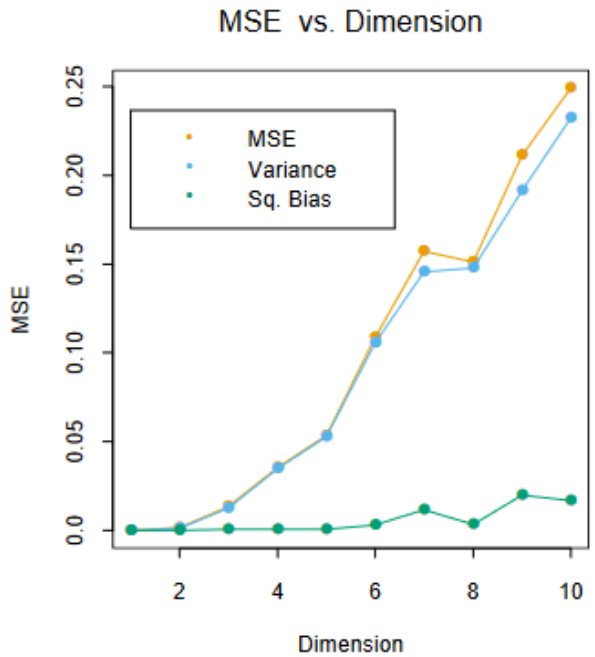
$\ell_i(x_0)$ - i-ty element $X(X^TX)^{-1}x_0$.

$$\begin{aligned}
\mathbb{E}PE(x_0) &= \mathbb{E}_{y_0|x_0}\mathbb{E}_\tau(y_0 - \hat{y}_0)^2 = \\
&= Var(y_0|x_0) + \mathbb{E}_\tau[\hat{y}_0 - \mathbb{E}_\tau\hat{y}_0]^2 + [\mathbb{E}_\tau\hat{y}_0 - x_0^T\beta]^2 = \\
&= Var(y_0|x_0) + Var_\tau(\hat{y}_0) + Bias^2(\hat{y}_0) = \\
&= \sigma^2 + \mathbb{E}_\tau x_0^T(X^TX)^{-1}x_0\sigma^2 + 0^2.
\end{aligned}$$

If N - large, τ - selected at random, $\mathbb{E}(X) = 0$, then $X^T X \to NCov(X)$ and

$$\begin{aligned}\mathbb{E}_{x_0}\mathbb{E}PE(x_0) &\approx \mathbb{E}x_0 x_0^T Cov(X)^{-1} x_0 \sigma^2/N + \sigma^2 = \\ &= trace[Cov(X)^{-1} Cov(x_0)]\sigma^2/N + \sigma^2 = \\ &= \sigma^2(p/N) + \sigma^2.\end{aligned}$$


1-NN in One Dimension
f(X)
X
0
1
2
3
4
-1.0
-0.5
0.0
0.5
1.0


MSE vs. Dimension
MSE
Dimension
MSE
Variance
Sq. Bias
0.0
0.05
0.10
0.15
0.20
0.25
2
4
6
8
10