1 Training data in supervised learning

Training data: $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, given/known.
Inputs $x_i \in \mathcal{X}$, outputs $y_i \in \dagger$.
For each example, what is $\mathcal{X}, \mathcal{Y}$?

- Image classification
- Image segmntation
- Translation
- Spam filtering

Usually $\mathcal{X} = \mathbb{R}^p$ (e.g. word counts in emails).
Notation.

- Inputs $x \in \mathbb{R}^p$.
- p = dimension of each input.
- n = number of training examples.
- Outputs $y \in \mathbb{R}$ for regression, $y \in \{0, 1\}$ for binary classification.
- $f : \mathcal{X} \to \mathcal{Y}$ is the prediction function we want to learn.
- Regression function $f : \mathbb{R}^p \to \mathbb{R}$
- Binary classification function $f : \mathbb{R}^p \to \{0, 1\}$.

Geometric interpretation of regression, height son/father (linear pattern), species versus temperature (non-linear pattern).
Draw residual r_i on graphs as vertical line segments.

2 Test data

Test data: $D' = \{(x'_1, y'_1), \ldots, (x'_m, y'_m)\}$, unknown.
Goal is generalization to new/test data: algo $\textsc{Learn}(D) = f$, with $f(x'_i) \approx y'_i$ for all test data, i.e. minimize

$$\text{Err}_{D'}(f) = \sum_{(x', y') \in D'} \ell[f(x'), y']$$

, where ℓ is a loss function:

- In binary classification we typically use the mis-classification-rate/0-1-loss $\ell[f(x), y] = I[f(x)! = y]$.
- In regression we use the squared error $\ell[f(x), y] = [f(x) - y]^2$.

Since D' is unknown, we need to assume that $D, D' \sim \mathcal{P}$.

3 Nearest neighbors algorithm

What is near? $\textsc{Dist} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is a non-negative distance function between two data points.
e.g. for $\mathcal{X} = \mathbb{R}^p$ we use

- L1 distance $\textsc{Dist}(x, x') = ||x - x'||_1 = \sum_{j=1}^{p} |x_j - x'_j|$
- L2 distance $\textsc{Dist}(x, x') = ||x - x'||_2 = \sqrt{\sum_{j=1}^{p} (x_j - x'_j)^2}$

Compute distances $d_1, \ldots, d_n \in \mathbb{R}_+$ where $d_i = \textsc{Dist}(x_i, x)$.
Example graph, draw horizontal line segments d_i between x, x_i.
Define neighbors function $N_{D,K}(x') = \{t_1, t_2, \ldots, t_K\}$, where
$t_1, \ldots, t_n \in \{1, \ldots, n\}$ are indices such that
distances $d_{t_1} \leq \cdots \leq d_{t_n}$ are sorted ascending.
Predict the mean output value of the K nearest neighbors,

$$f_{D,K}(x') = \frac{1}{K} \sum_{i \in N_{D,K}(x')} y_i$$

- For regression this is the mean.
- For binary classification the mean is intepreted as a probability, predict 1 if greater than 0.5, and predict 0 otherwise.

TODO: Geometric interpretation of binary classification in 1d,2d.