K-NEAREST NEIGHBOURS

LECTURE 2
SECTION 2
JUNE 2ND

IACS
INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

UNIVERSITY of
RWANDA

MOTIVATION

MODEL INTERPRETABILITY:

Interpreting our models help us evaluate them and extract new insights about our data.

y = price (\$)

X_1 = size (m^2)

X_2 = Distance to city center (m)

M1: Linear Regression

$$y = 3.5X_1 - 0.1X_2 + 100{,}000$$

Train MSE: 1,200
Test MSE: 2,200

Is this a reasonable model?

What does it say about housing prices?

M2: Polynomial Regression

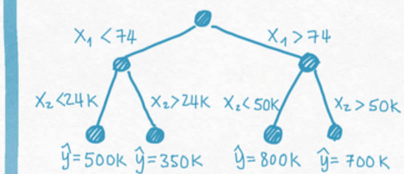
$$y = 0.01X_1^2 - 0.02X_2^2 + 2.1X_1X_2 + 0.1X_1 - 0.2X_2 + 10{,}000$$

Train MSE: 358
Test MSE: 657

Is this a reasonable model?

What does it say about housing prices?

M3: Regression Tree



Train MSE: 568
Test MSE: 1,000

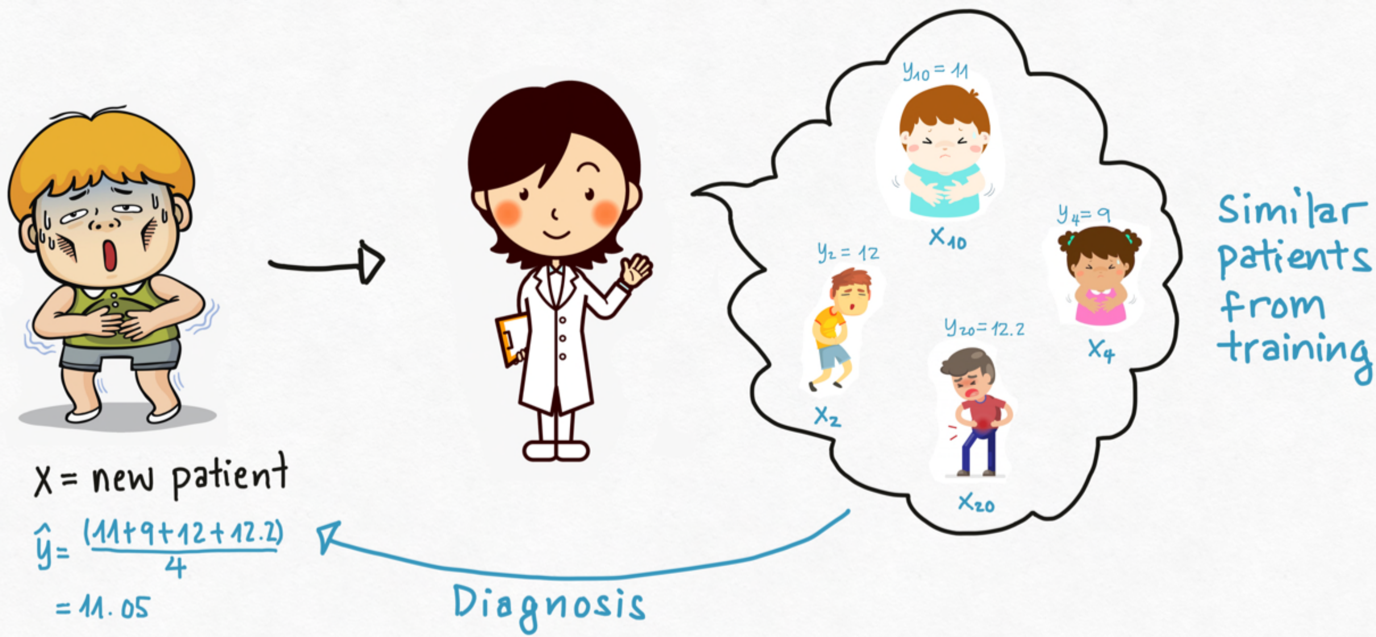
Is this a reasonable model?

What does it say about housing prices?

Which model should we choose?

EXPLANATION BY EXAMPLE:

Linear models and regression trees are interpretable in different ways. Another way to explain or interpret decisions is by looking at examples:



K-NEAREST NEIGHBOURS

K-NEAREST NEIGHBOURS:

The very human way of decision making by similar examples can be formalized as an algorithm:

The k-Nearest Neighbor Algorithm:

Given a data set $D = \{(X^{(1)}, y^{(1)}), \ldots (X^{(N)}, y^{(N)})\}$. For every new X:

1. Find k-number of observations in D most similar to X:

$$\{(X^{(n_1)}, y^{(n_1)}), \ldots, (X^{(n_k)}, y^{(n_k)})\}$$

These are called the k-nearest neighbours of x.

2. Average the output of the k-nearest neighbours of x:

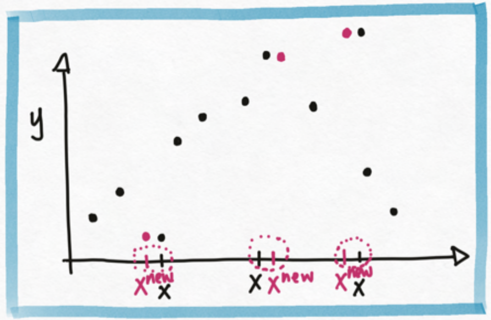
$$\hat{y} = \frac{1}{K} \sum_{k=1}^{K} y^{(n_k)}$$

3. Predict $\hat{y}$ for x.

OVERFITTING AND UNDERFITTING:

- Training Data
- prediction for new data

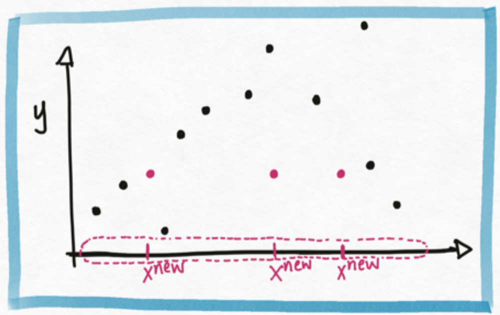
K=1



predictions on new data is influenced by noise in the data.

test error will be high.

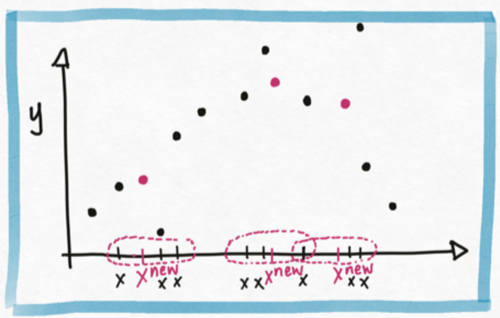
K=10



predictions is unaffected by local variations in the data.

test error will be high.

K=3



predictions take into account local variations but averaging several neighbours cancels the effect of noise.

COMPARING MODELS

COMPARISON OF MODELS:

Choosing the right model isn't just about minimizing test error. We want to extract insights from our models.

	probabilistic	Has a fixed form $f_w(x)$ (parametric)	Easy to interpret
Linear Regression	YES	YES	YES
Polynomial Regression	YES	YES	NO
Regression Tree	NO	YES	If the tree is not big
K-Nearest Neighbours	NO	NO	YES
	Explicitly modeling the noise as a RV help us diagnose & improve the model.	Having an explicit functional form $f_w(x)$ makes it easy to store and use the model.	Interpretation helps us evaluate our model and understand the data.