Supervised learning

Classification

Unsupervised learning

Classification :

dataset X

$(\vec{x}_i, y_i)$

$\vec{x}_i$ — feature vector

y_i — label

total error rate: % of attempts that are wrong

accuracy: % attempts correct

2-class

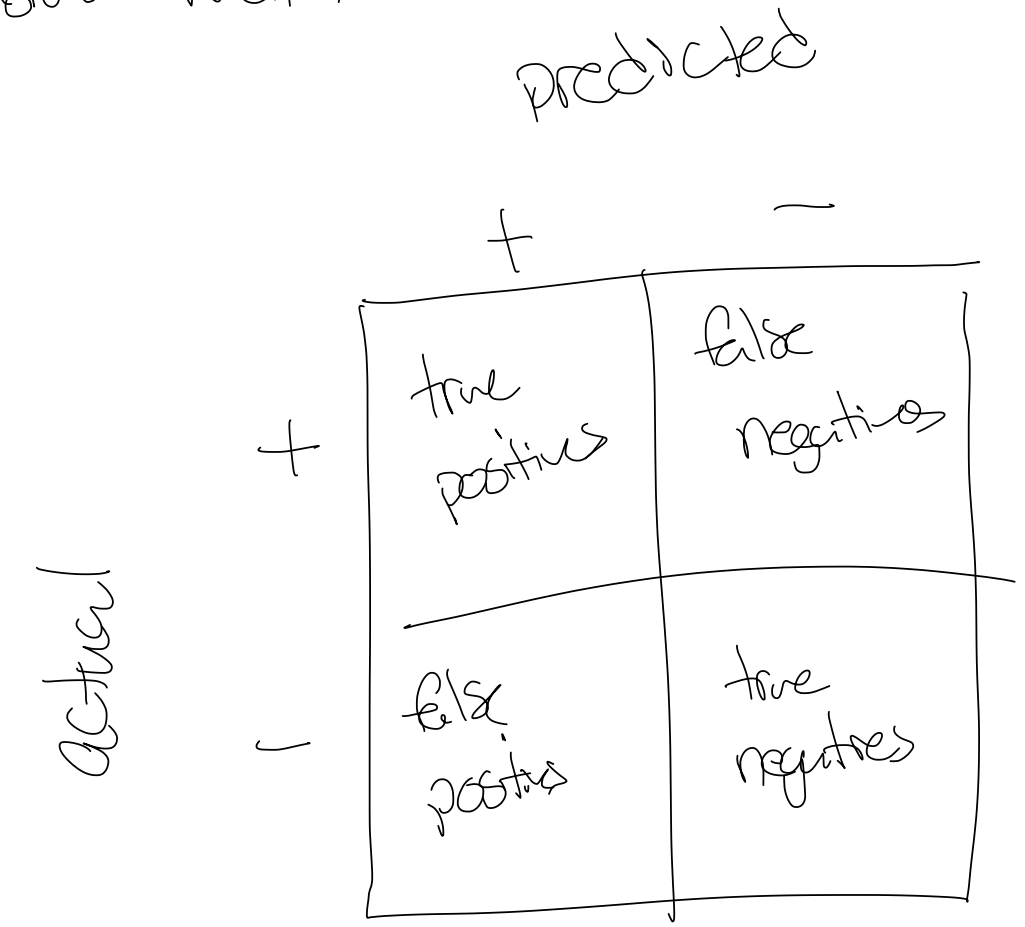
false positive rate: % of − classified as +

false neg. rate: % of + classified as −

sensitivity: % of true + classified as +

specificity: % of true − classified as −

Confusion matrix



predicted

actual	0	1	2	class error rate
0	100	6	4	$10/110 = 9.09\%$
1	20	4	9	$29/33$
2	5	2	0	100%

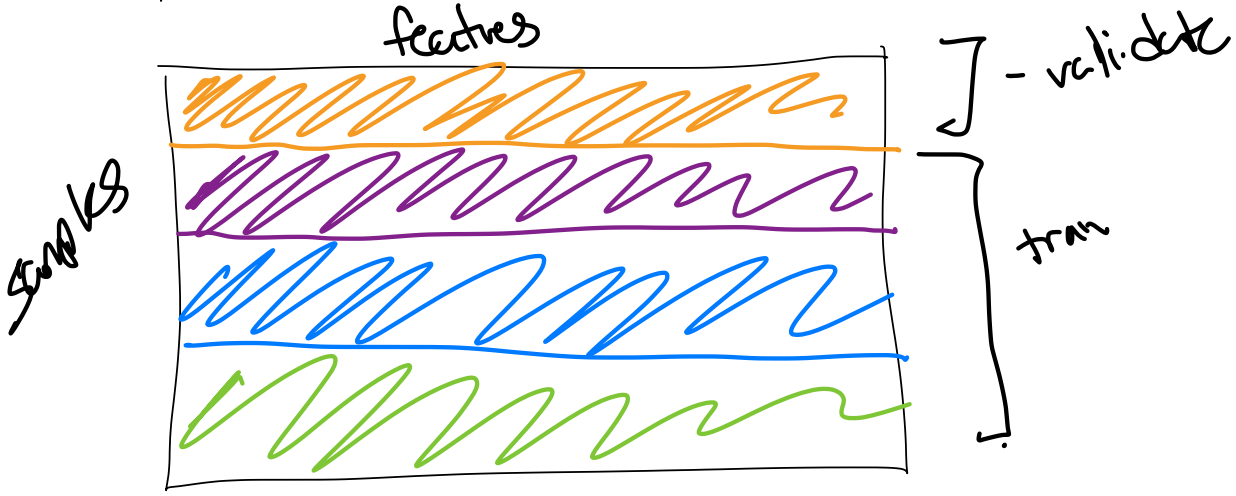
training data

testing data 15 - 20%

Validation

60/20/20 or 70/15/15

K-fold cross-validation



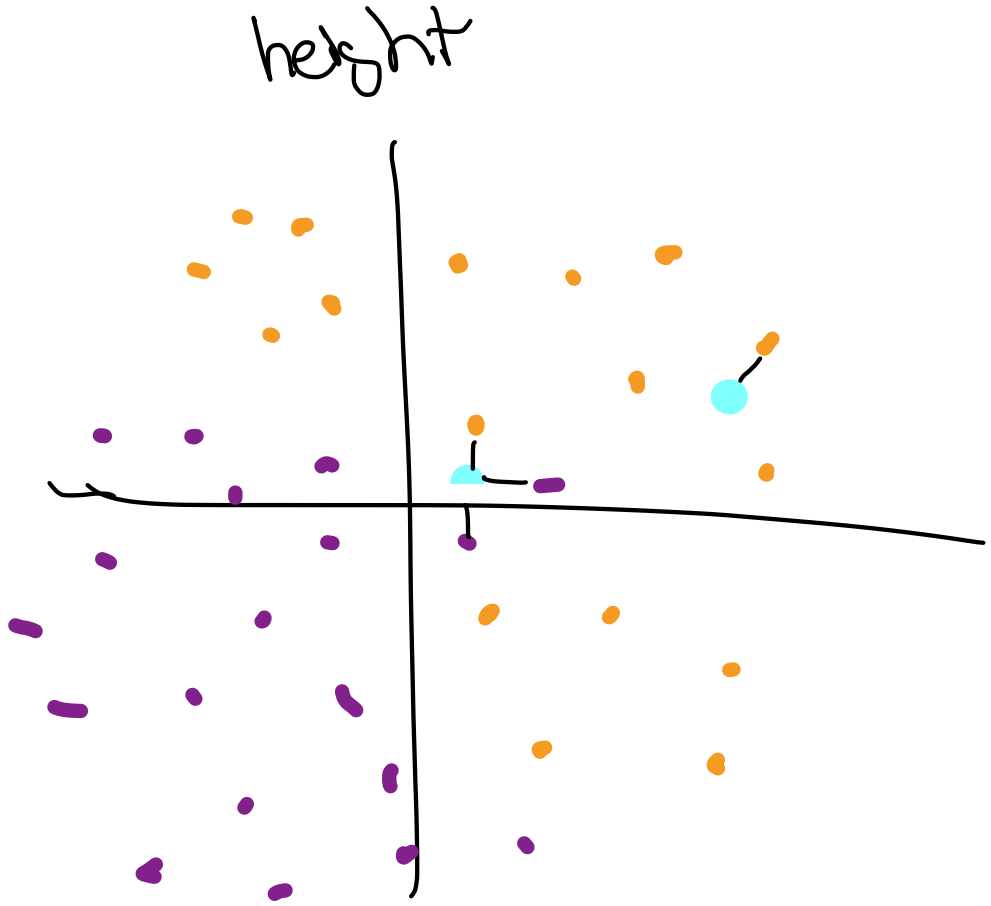
average all

↓

choose model

Nearest Neighbors

$(\vec{x}_i, y_i)$



$\vec{x}$

$\vec{v}$ $\vec{w}$

if $\vec{v}$ and $\vec{w}$ are close

then $p(y|\vec{v})$ and $p(y|\vec{w})$ are close

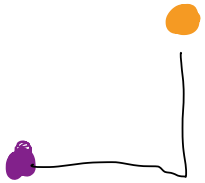
1) a lot of data

2) when are points close

euclidean distance

$$\sqrt{\sum_{i=1}^{d} (v_i - u_i)^2} = \|v - u\|_2$$

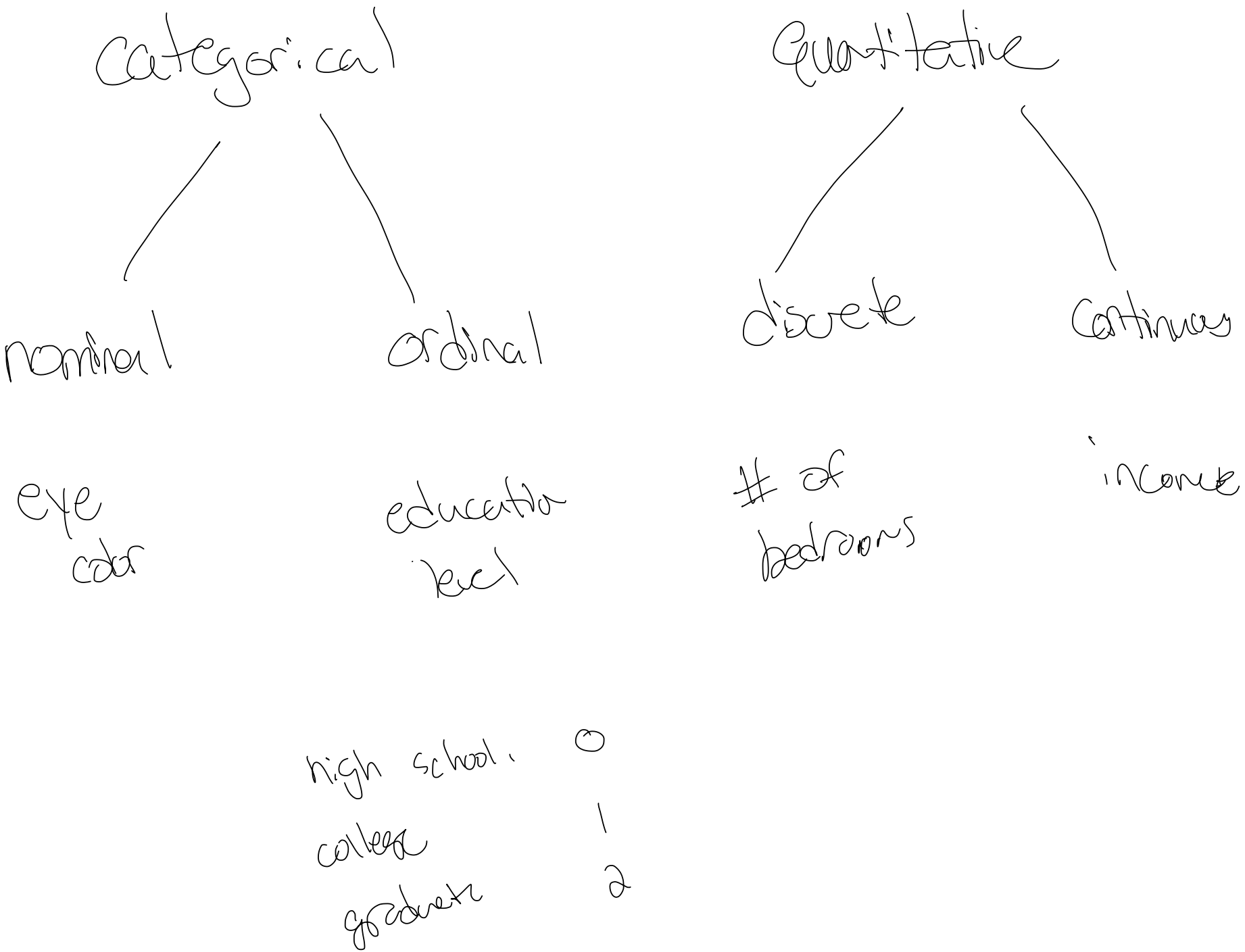
manhatten distance



$$\sum_{i=1}^{d} |v_i - u_i| = \|v - u\|_1$$

minkowski distance

$$\left(\sum_{i=1}^{d} |v_i - u_i|^p\right)^{1/p} = \|v - u\|_p$$


Categorical
Quantitative
nominal
ordinal
discrete
Continuous
eye color
education level
# of bedrooms
income
high school. 0
college 1
graduate 2

hamming distance

$$\frac{\text{num\_unequal}}{\text{num\_total}}$$

$[0, 1]$

$[3, 1000]$

$$\frac{v - v_{min}}{v_{max} - v_{min}}$$

corrected from class