KNN for Categorical Features

Dr. Muhammad Wasim

Dataset Example

Color	HWY MPG	Car Type
White	23	Van
Red	28	Sport
Black	32	Sport
Red	42	Sedan
Red	40	Sedan
White	20	Van

$$\delta(val_i, val_j) = \sum_{h=1}^{\#\ of\ classes} \left| \mathrm{P}(\mathrm{c_h}|\mathrm{val_i}) - \mathrm{P}(\mathrm{c_h}|\mathrm{val_j}) \right|^2$$

Value Difference Metric (VDM)

- VDM was introduced in 1986 to provide an appropriate distance function for symbolic attributes.
- It's based on the idea that the goal of finding the distance is to find the right class. One way to measure this is to look at the following conditional probabilities.

- $\delta(val_i, val_j) = \sum_{h=1}^{\#\ of\ classes} | \mathrm{P}(\mathrm{c_h}|\mathrm{val_i}) - \mathrm{P}(\mathrm{c_h}|\mathrm{val_j}) |^2$
- You can then plug the delta in the Euclidean Distance:

$$d(P, Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Solution

- $\delta(val_i, val_j) = \sum_{h=1}^{\#\ of\ classes} | P(c_h|val_i) - P(c_h|val_j) |^2$

Color	HWY MPG	Car Type
White	23	Van
Red	28	Sport
Black	32	Sport
Red	42	Sedan
Red	40	Sedan
White	20	Van

Class 1:Van	Class 2: Sport	Class 3: Sedan
P(Van|White)	P(Sport|White)	P(Sedan|White)
P(Van|Red)	P(Sport|Red)	P(Sedan|Red)
P(Van|Black)	P(Sport|Black)	P(Sedan|Black)

Solution (Cont.)

Color	Car Type
White	Van
Red	Sport
Black	Sport
Red	Sedan
Red	Sedan
White	Van

Class 1:Van	Class 2: Sport	Class 3: Sedan
P(Van | White) 1	P(Sport | White) 0	P(Sedan | White) 0
P(Van | Red) 0	P(Sport | Red) 1/3	P(Sedan | Red) 2/3
P(Van | Black) 0	P(Sport | Black) 1	P(Sedan | Black) 0

Solution (Cont.)

Class 1:Van	Class 2: Sport	Class 3: Sedan
P(Van|White) 1	P(Sport|White) 0	P(Sedan|White) 0
P(Van|Red) 0	P(Sport|Red) 1/3	P(Sedan|Red) 2/3
P(Van|Black) 0	P(Sport|Black) 1	P(Sedan|Black) 0

$$\delta(val_i, val_j) = \sum_{h=1}^{\#\ classes} |P(c_h|val_i) - P(c_h|val_j)|^2$$

$\delta(Red, White) = |P(Van|Red) - P(Van|White)|^2 + |P(Sport|Red) - P(Sport|White)|^2 + |P(Sedan|Red) - (Sedan|White)|^2$

$\delta(Red, White) = (0-1)^2 + (1/3 - 0)^2 + (2/3-0)^2 = 1.556$

$\delta(Red, Black) = |P(Van|Red) - P(Van|Black)|2 + |P(Sport|Red) - P(Sport|Black)|^2 + |P(Sedan|Red) - (Sedan|Black)|^2$

$\delta(Red, Black) = (0-0)^2 + (1/3 - 1)^2 + (2/3-0)^2 = 0.889$

$\delta(Red, Red) = 0$

Solution (cont.)

Color	HWY MPG	Car Type	Distance
White	23	Van	$\sqrt{1.55 + (28 - 23)^2}$=5.153
Red	28	Sport	$\sqrt{0 + (28 - 28)^2}$=0
Black	32	Sport	$\sqrt{0.889 + (28 - 32)^2}$=4.110
Red	42	Sedan	$\sqrt{0 + (28 - 42)^2}$=14
Red	40	Sedan	$\sqrt{0 + (28 - 40)^2}$=12
White	20	Van	$\sqrt{1.55 + (28 - 20)^2}$=8.096
Red	28	?	

- Using 3-NN, the prediction will be that her car is a sport car.