

Nearest Neighbor Classifiers

Hal Daumé III

CS5350: Machine Learning

22 January 2008

1. The geometric view of data
2. Dealing with categorical values
3. One-nearest-neighbor classification
4. K -nearest-neighbor classification
5. Extensions...

The geometric view of data

We've already talked about representing a data point as a vector:

- ▶ The **length** of the vector is the total number of features
- ▶ Each **element** in the vector is the associated feature value

Once we've done this, we can think about data as points in high-dimensional space.

This enables us to use tools from **linear algebra** to think about learning problems:

- ▶ Distance between examples \Leftarrow [this lecture](#)
- ▶ Linear transformations
- ▶ Projections onto subspaces

Reminder: Euclidean distance

Given $\mathbf{x} = \langle x_1, x_2, \dots, x_D \rangle$ and $\mathbf{y} = \langle y_1, y_2, \dots, y_D \rangle$, the **Euclidean distance** between \mathbf{x} and \mathbf{y} is defined by:

$$\begin{aligned}
 \|\mathbf{x} - \mathbf{y}\| &= \sqrt{\sum_{d=1}^D (x_d - y_d)^2} \\
 &= \sqrt{\sum_{d=1}^D (x_d^2 + y_d^2 - 2x_d y_d)} \\
 &= \sqrt{\sum_{d=1}^D x_d^2 + \sum_{d=1}^D y_d^2 - 2 \sum_{d=1}^D x_d y_d} \\
 &= \sqrt{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\mathbf{x} \cdot \mathbf{y}}
 \end{aligned}$$

This confirms that dot products are measures of similarity

Embedding the play/no-play data

Y	Out	T	R		Y	$\langle \text{Out}, \text{T}, \text{R} \rangle$
P	Sunny	Low	Yes	⇒	1	$\langle ? , 0 , 1 \rangle$
N	Sunny	High	Yes		0	$\langle ? , 1 , 1 \rangle$
N	Sunny	High	No		0	$\langle ? , 1 , 0 \rangle$
P	Overcast	Low	Yes		1	$\langle ? , 0 , 1 \rangle$
P	Overcast	High	No		1	$\langle ? , 1 , 0 \rangle$
P	Overcast	Low	No		1	$\langle ? , 0 , 0 \rangle$
N	Rainy	Low	Yes		0	$\langle ? , 0 , 1 \rangle$
P	Rainy	Low	No		1	$\langle ? , 0 , 0 \rangle$

Why not just map “Sunny” to 0, “Overcast” to 1 and “Rainy” to 2?

Dealing with categorical values

Solution: map a categorical feature with K values into K binary features.

Y	Out	T	R		Y	$\langle \text{S?}, \text{O?}, \text{R?}, \text{T}, \text{R} \rangle$
P	Sunny	Low	Yes	⇒	1	$\langle 1 , 0 , 0 , 0 , 1 \rangle$
N	Sunny	High	Yes		0	$\langle 1 , 0 , 0 , 1 , 1 \rangle$
N	Sunny	High	No		0	$\langle 1 , 0 , 0 , 1 , 0 \rangle$
P	Overcast	Low	Yes		1	$\langle 0 , 1 , 0 , 0 , 1 \rangle$
P	Overcast	High	No		1	$\langle 0 , 1 , 0 , 1 , 0 \rangle$
P	Overcast	Low	No		1	$\langle 0 , 1 , 0 , 0 , 0 \rangle$
N	Rainy	Low	Yes		0	$\langle 0 , 0 , 1 , 0 , 1 \rangle$
P	Rainy	Low	No		1	$\langle 0 , 0 , 1 , 0 , 0 \rangle$