K Nearest Neighbor

CMPUT 328

Nilanjan Ray

Computing Science, University of Alberta, Canada

Supervised machine learning: the tabular view

Independent variable (aka features or predictors)

Output / Prediction / dependent variable

x_1	x_2	x_3	x_4	y
1.2	-3.9	4.0	0	1.6
2.1	2.4	-0.7	-0.2	1.2
…	…	…	…	…
…	…	…	…	…
3.2	…	…	1.9	0.3
1.4	…	…	1.5	?
3.1	…	…	2.1	?

Training data: complete table

Test data: incomplete table

ML learns to map **x** to y

In other words, ML learns a function, *f* so that $y = f(\mathbf{x})$

The function *f* is called prediction function

K nearest neighbor (K-nn)

- K-nn uses a very simple form of learning: it remembers all of *m* training data points, i.e., $\{(x_i^{tr}, y_i^{tr})\}_{i=1}^{m}$!
- Note: x_i^{tr} is a vector of dimension 1-by-*d*, so $x_i^{tr} = [x_{i,1}^{tr}, x_{i,2}^{tr}, \ldots, x_{i,d}^{tr}]$
- To predict the output for a test data point x, k-nn computes *k* nearest training examples to *x*: $\{(x_i^{tr}, y_i^{tr})\}_{i \in N_k(x)}$, where $N_k(x)$ is the set of indices of *k* training data points that are closest to *x*.
- The output for x is computed by aggregating *k* responses: $f(x) = Ave_{i \in N_k(x)}(y_i)$.
- Using a validation set, we find out the right value of *k*.

K-nn: A toy example

Training data, $m = 5$

Test data point

x_1	x_2	y
2	-1	0
3	2	1
0	4	0
-2	5	0
2	0	1
1	1	?

For this problem, note that the feature vector dimension, d=2

Let's assume $k = 3$

To find out k=3 nearest neighbors, compute distances:

$D_1([1, 1], [2, -1]) = |1-2|+|1+1| = 3$

$D_2([1, 1], [3, 2]) = |1-3|+|1-2| = 3$

$D_3([1, 1], [0, 4]) = |1-0|+|1-4| = 4$

$D_4([1, 1], [-2, 5]) = |1+2|+|1-5| = 7$

$D_5([1, 1], [2, 0]) = |1-2|+|1-0| = 2$

So, k=3 nearest neighbors are

$N_3([1,1]) = \{1, 2, 5\}$

Prediction for test data point:

$f([1, 1]) = Ave([y(1), y(2), y(5)])$

$= Ave([0, 1, 1]) = 1$

Here, we computed "Ave" by taking mode.

How to find the right value of *k*?

- Divide training data into two sets: training (90%) and validation (10%).
- For each *k* in a range, find out k-nn prediction accuracy on the validation set.
- Choose the *k* that has yielded the highest accuracy on the validation set.

MNIST digit image classification

Small 28 pixels-by-28 pixels images of hand written digits

The visual recognition problem definition:
to recognize the digit from an image

We can attempt to solve this using k-nn.

Feature dimension, d = 28 * 28 = 784

Pixel values (feature) | Digit

	x_1	x_2	…	x_{784}	y
Training data	0.1	0.3	…	0.0	0
	0.2	0.1	…	0.5	1
	…	…	…	…	…
	…	…	…	…	…
	0.0	0.98	…	0.8	9
Test data	0.5	0.25	…	0.36	?
	0.1	0.95	…	0.1	?

A recommended resource:
https://cs231n.github.io/classification/