1 K-Nearest Neighbors

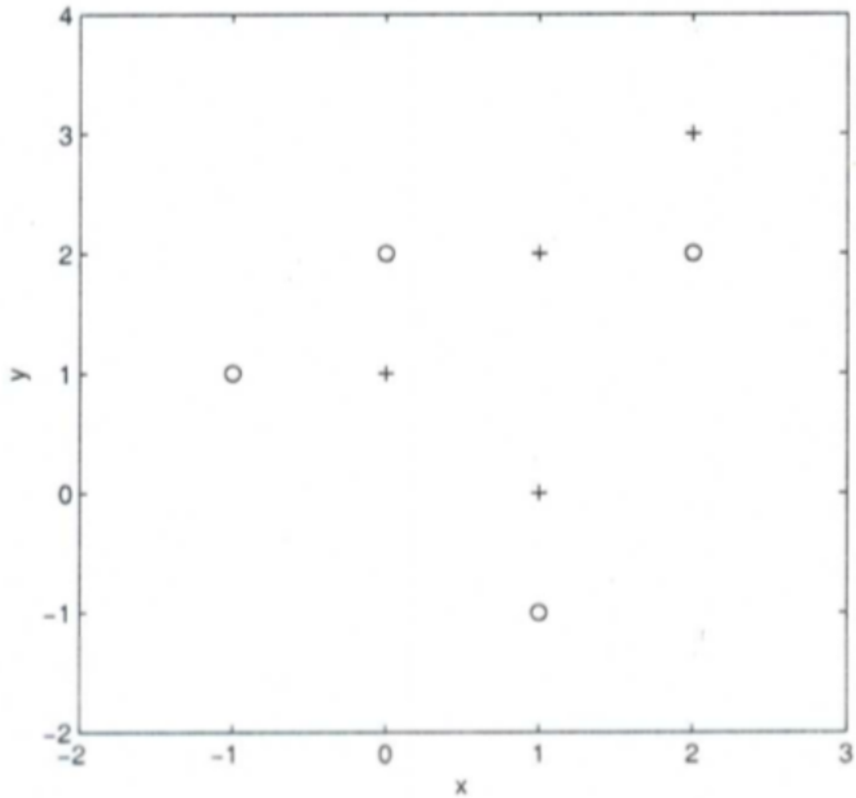
K-nearest neighbors is a nonparametric model that, given a point, predicts the mode of the classes of the k nearest points.

1.1 KNN Example

Consider the following training set in the 2-dimensional Euclidian space:

x	y	Class
-1	1	$-$
0	1	$+$
0	2	$-$
1	-1	$-$
1	0	$+$
1	2	$+$
2	2	$-$
2	3	$+$

The figure below shows a visualization of the data.



1. What is the prediction of the 3-nearest-neighbor classifier at the point $(1, 1)$?

   $+$

2. What is the prediction of the 5-nearest-neighbor classifier at the point $(1, 1)$?

   $+$

3. What is the prediction of the 7-nearest-neighbor classifier at the point $(1, 1)$?

   $-$

1.2 Conceptual Questions

Do smaller or larger values of k cause overfitting?

Smaller values of k cause overfitting, because if we predict with a smaller number of points we are more dependent on the individual data points in the training data.

If $k = 1$, what will we predict for a given point? What could be a problem with this?

We will predict the class of the nearest point. This could result in overfitting because we are too dependent on a single data point.

If $k = n$, where n is the number of data points, what will we predict for every point? What could be a problem with this?

We will predict the most common class across the entire data set, regardless of position. This is a problem because we aren't actually using the location of the point to change our prediction.

We see that both too large and too small k can lead to problems. See the powerpoint for more information about this, and more practice and graphs relating to k nearest neighbors.

2 Decision Trees

2.1 Review

In this recitation we will be going through a decision tree problem with greedy search. In order to understand greedy search, we will go over the ideas of entropy, conditional entropy, and mutual information.

Entropy is a measurement of the uncertainty in a random variable. We quantify this by asking, "On average, how many bits do we need to represent a single draw of this random variable?" Its formula is as follows:

$$H(Y) = -\sum_y P(Y = y) \lg P(Y = y)$$

Let's look at two simple examples. Let Y have two values, A and B.

If $P(Y = A) = 1$, a random draw of Y doesn't give us any additional information - it will always be A. We see that

$$H(Y) = -P(Y = A) \lg P(Y = A) - P(Y = B) \lg P(Y = B) = -1 \lg 1 - 0 \lg 0 = 0.$$

So entropy in this case is 0, which makes sense because we don't need any bits to represent which value Y took - it's always A.

If $P(Y = B) = \frac{1}{2}$, then there is an equal chance for both values of Y, so we are the least confident about what Y will be. We see that

$$H(Y) = -P(Y = A) \lg P(Y = A) - P(Y = B) \lg P(Y = B) = -\frac{1}{2} \lg \frac{1}{2} - \frac{1}{2} \lg \frac{1}{2} = -\lg \frac{1}{2} = 1.$$

So entropy in this case is 1, which means we need 1 bit to store the value Y took. This makes sense because there are 2 values of Y, so we could use a single bit and use 0 to represent A and 1 to represent B.

Conditional entropy is the expected value of the entropy of Y given X, over all values of X. This lets us quantify the entropy of Y given that we know X.

$$H(Y|X) = \sum_x P(X = x) H(Y|X = x)$$

Mutual information is a measurement of the information we gain about Y by observing X. We get this by finding the difference between the entropy of Y, and the conditional entropy of Y given X:

$$I(Y; X) = H(Y) - H(Y|X)$$

If $I(Y; X)$ is large, then we gained a lot of information about Y by observing X. If $I(Y; X)$ is 0, then we did not gain any information about Y by observing X, so we know X and Y are independent.

2.2 Practice

Refer to the recitation slides posted for an example problem.