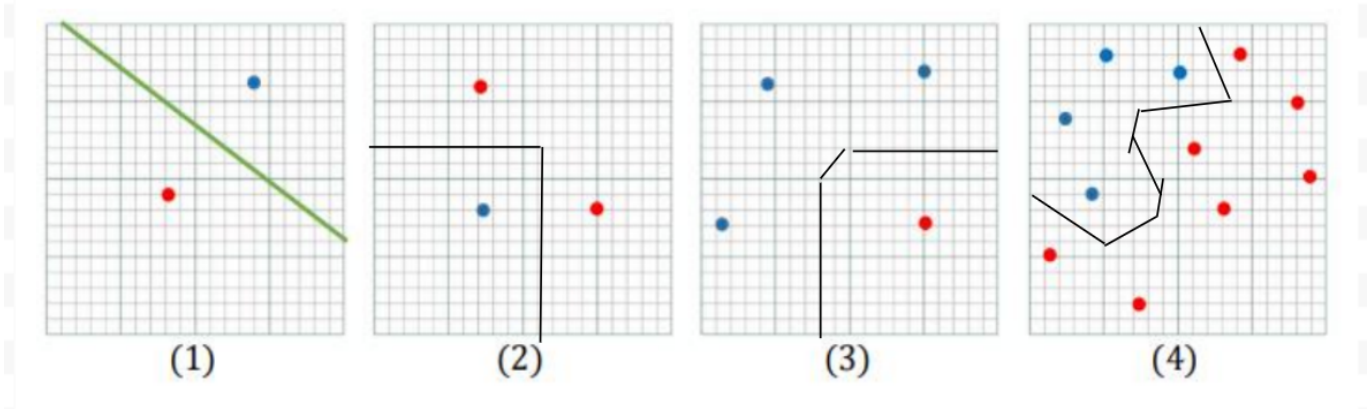
دانشگاه صنعتی اصفهان
دانشکده مهندسی برق و کامپیوتر

عنوان: تکلیف دوم درس داده کاوی

نام و نام خانوادگی: نیلوفر سعیدی
شماره دانشجویی: ۹۸۲۲۹۶۳

1

1.1



1.2

Yes, always. Because in this case, only the marginal records need to be stored instead of all of the records, taking up much less space.

1.3

No, we can always add new training data to KNN without the need of extra consideration. The reason is that KNN uses the available data and calculates the distances in each prediction, unlike decision trees that predict on an already built and stored model.

2

Decision trees normally grow until they to overfit on the data, meaning that they memorize every single record(including existing noise). By pruning the extra leaves, we prevent this from happening.

In pre-pruning, we set a limitation for the number of leaves, before the tree is constructed. In post-pruning, after the tree is thoroughly constructed, we cut or replace some sub-trees that we find unnecessary or trouble-making.

3

The question:

$$p(hasCold?|hasFever = 1, coughs = 0, hasHeadacke = 1)$$

We calculate the probability of having or not having a cold separately, given the above conditions. Then we compare these probabilities and the higher one will determine the class(yes or no) of the attribute(hasCold).

$p(hasHeadache = 1|hasCold = yes) = 2/5$
$p(coughs = 0|hasCold = yes) = 1/5$
$p(hasFever = 1|hasCold = yes) = 3/5$
$p(hasFever = 1, coughs = 0, hasHeadacke = 1|hasCold = yes) = 2/5 * 1/5 * 3/5 = 6/125$

$p(hasHeadache = 1|hasCold = no) = 3/5$
$p(coughs = 0|hasCold = no) = 3/5$
$p(hasFever = 1|hasCold = no) = 2/5$
$p(hasFever = 1, coughs = 0, hasHeadacke = 1|Y = hasCold = no) = 2/5 * 3/5 * 3/5 = 18/125$

Therefore, hasCold = no