Exercise collection – Supervised Regression

Lecture exercises

Exercise 1:

Let the 2D feature vectors in the following figure be with two different numeric target values (2 and 4). Predict the point (7,6) - represented by the grey point in the picture - with the k-nearest neighbor method. Distance function should be the L_1 norm (Manhattan distance):

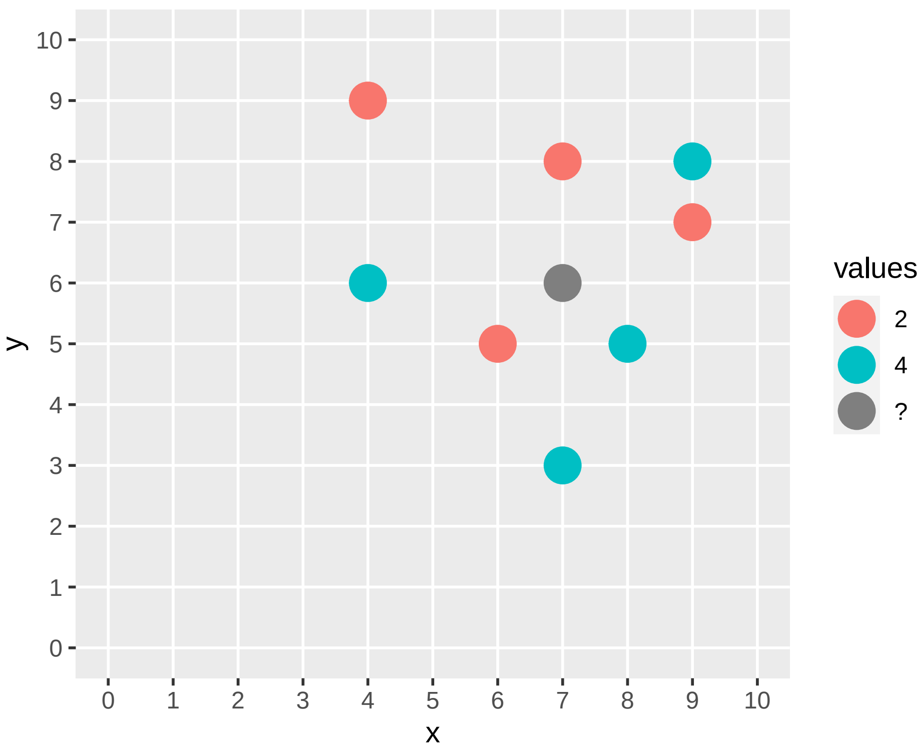
$$d_{\text{manhattan}}(x, \tilde{x}) = \sum_{j=1}^{p} |x_j - \tilde{x}_j|$$

State as the prediction the unweighted and the weighted (according to the Manhattan distance) mean of the values of the k-nearest neighbors.

a) $k = 3$

b) $k = 5$

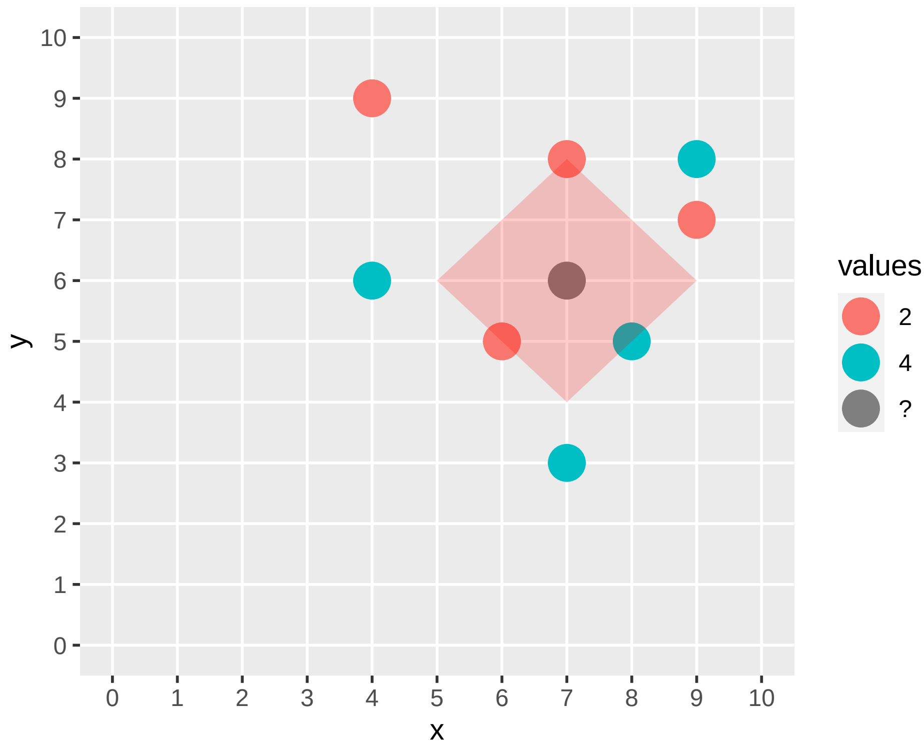
c) $k = 7$



Solution 1:

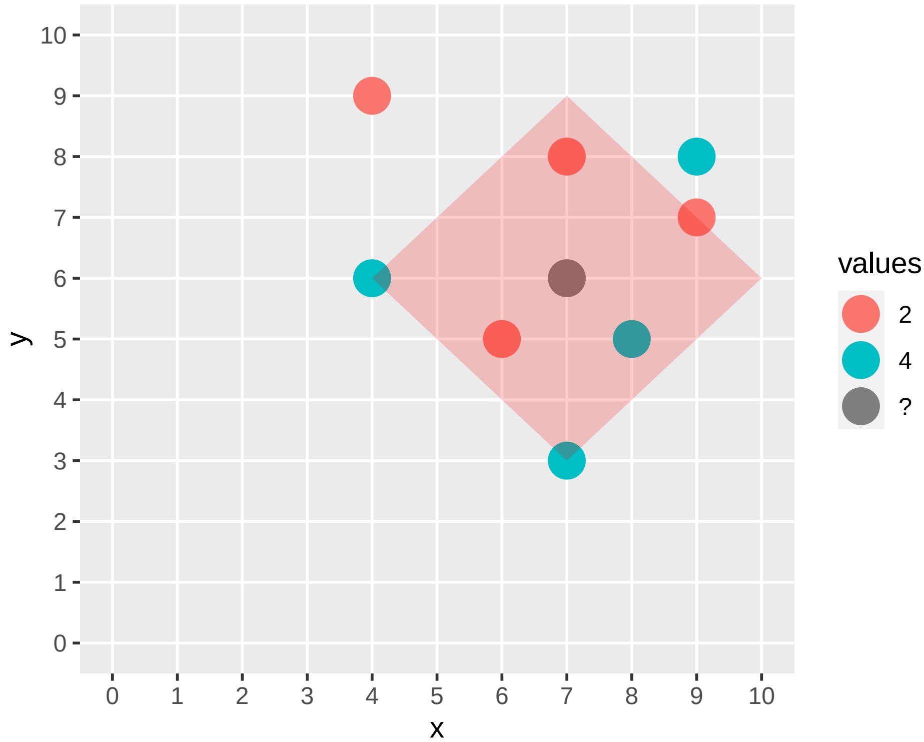
a) $k = 3$

$$\hat{y} = \frac{2+2+4}{3} = \frac{8}{3} \approx 2.67$$
$$\hat{y}_{\text{weighted}} = \frac{\frac{1}{2}\cdot 2 + \frac{1}{2}\cdot 2 + \frac{1}{2}\cdot 4}{\frac{3}{2}} = \frac{8}{3} \approx 2.67$$



b) $k = 5$

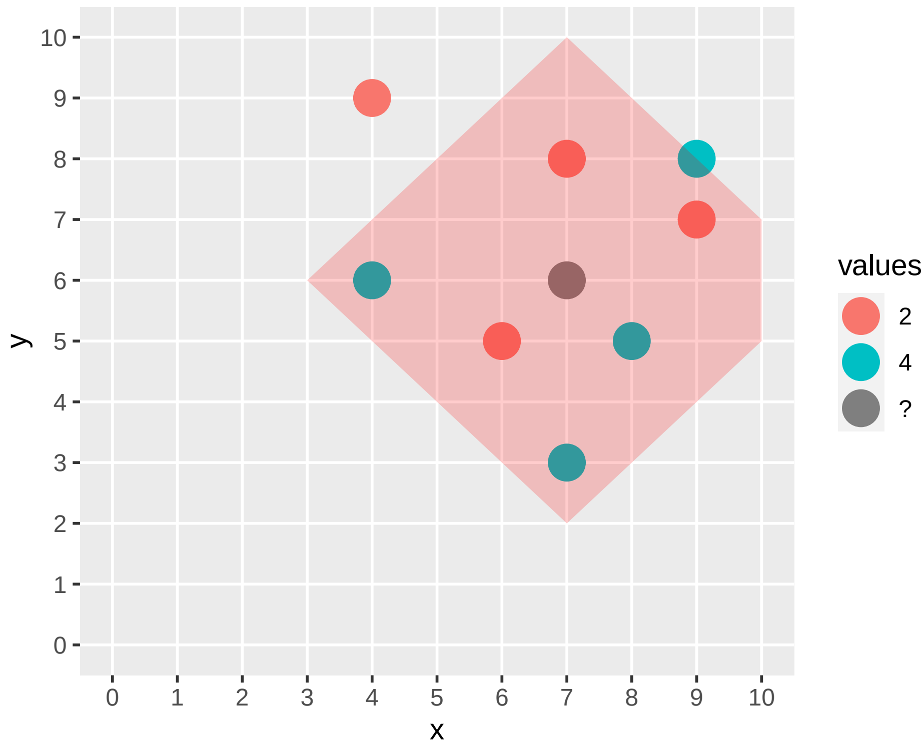
$$\hat{y} = \frac{2+2+2+4+4+4}{6} = 3$$
$$\hat{y}_{\text{weighted}} = \frac{\frac{1}{2}\cdot 2 + \frac{1}{2}\cdot 2 + \frac{1}{3}\cdot 2 + \frac{1}{2}\cdot 4 + \frac{1}{3}\cdot 4 + \frac{1}{3}\cdot 4}{\frac{5}{2}} = \frac{44}{15} \approx 2.93$$



c) $k = 7$

$$\hat{y} = \frac{2+2+2+4+4+4+4}{7} = \frac{22}{7} \approx 3.14$$

$$\hat{y}_{\text{weighted}} = \frac{\frac{1}{2}\cdot 2 + \frac{1}{2}\cdot 2 + \frac{1}{3}\cdot 2 + \frac{1}{2}\cdot 4 + \frac{1}{3}\cdot 4 + \frac{1}{3}\cdot 4 + \frac{1}{4}\cdot 4}{\frac{11}{4}} = \frac{100}{33} \approx 3.03$$



Exercise 2:

How in mlr3 a learner can be constructed and what it represents can be found at `https://mlr3book.mlr-org.com/learners.html`.

a) How does a learner in mlr3 compare to what you've learned in the videos?

b) Pick an mlr3 learner of your choice. What are the different settings for this learner?
(Hint: Use `mlr_learners$keys()` to see all available learners)

Solution 2:

a) Learning consists of *representation* (hypothesis space), *evaluation* (risk) and *optimization*.
A learner in mlr3 can be thought of as the implementation of these components, since

- a representation of the associated model learnt from the data by using the implemented optimization is stored in such a learner object,
- its performance measures can be accessed afterwards.

b)
```r
library(mlr3)
library(mlr3learners)

# show all available learners
mlr_learners$keys()

##  [1] "classif.cv_glmnet"   "classif.debug"       "classif.featureless"
##  [4] "classif.glmnet"      "classif.kknn"        "classif.lda"
##  [7] "classif.log_reg"     "classif.multinom"    "classif.naive_bayes"
## [10] "classif.nnet"        "classif.qda"         "classif.ranger"
## [13] "classif.rpart"       "classif.svm"         "classif.xgboost"
## [16] "regr.cv_glmnet"      "regr.featureless"    "regr.glmnet"
## [19] "regr.kknn"           "regr.km"             "regr.lm"
## [22] "regr.ranger"         "regr.rpart"          "regr.svm"
## [25] "regr.xgboost"        "surv.cv_glmnet"      "surv.glmnet"
## [28] "surv.ranger"         "surv.xgboost"

# see settings for a specific learner, e.g., for a regression tree
rpart_learner <- lrn("regr.rpart")
print(rpart_learner)

## <LearnerRegrRpart:regr.rpart>
## * Model: -
## * Parameters: xval=0
## * Packages: rpart
## * Predict Type: response
## * Feature types: logical, integer, numeric, factor, ordered
## * Properties: importance, missings, selected_features, weights
```

Exercise 3:

We want to predict the age of an abalone using its longest shell measurement and its weight.
See: `http://archive.ics.uci.edu/ml/datasets/Abalone` for more details.

a) Plot `LongestShell`, `WholeWeight` on the x- and y-axis and color points with `Rings`

Using the mlr3-package:

b) Fit a linear model

c) Fit a k-nearest-neighbors model

d) Compare the fitted and observed targets for lm and knn, respectively (Hint: Use `autoplot()`)

Hint: See the official book manual of the mlr3 package for usage:

`https://mlr3book.mlr-org.com/index.html`

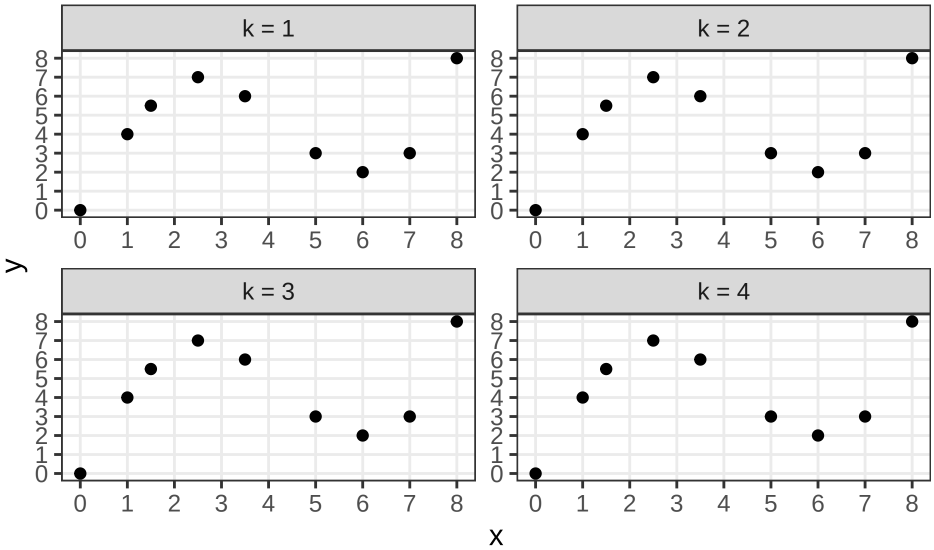
Solution 3:

See R code

Questions from past exams

Exercise 4: WS2020/21, retry exam, question 1

ID	$\mathbf{x}$	y
1	0.0	0.0
2	1.0	4.0
3	1.5	5.5
4	2.5	7.0
5	3.5	6.0
6	5.0	3.0
7	6.0	2.0
8	7.0	3.0
9	8.0	8.0



(a) Assume we use k-nearest neighbours regression with the L1 norm (Manhattan distance) as a distance function. For every $k \in \{1, 2, 3, 4\}$, do the following:

- (i) Mark the k-nearest neighbours of a new observation $\mathbf{x} = 4$ in the graphic below.
- (ii) Calculate the predicted value $\hat{y}$ for $\mathbf{x} = 4$ as the unweighted mean of the k-nearest neighbours and draw it in the graphic below.

(b) Would using the euclidean distance as distance measure in a) have made a difference? Explain your answer.

Solution 4:

(a) $\{x_5\}$, $\{x_5, x_6\}$, $\{x_4, x_5, x_6\}$, $\{x_4, x_5, x_6, x_7\}$

(b) No. In the case of a single feature L1 and L2 are identical.

Ideas & exercises from other sources