Computer Science Department, University of Bonn

Knowledge Graph Analysis

Solutions to Exercise Sheet 3

Dr. Hamed Shariat Yazdi, Prof. Jens Lehmann

November 27, 2018

1 In Class

1. **Classification and Regression**

   a) **1.** X consists of images of fruit, and Y = {apple, orange, pear, banana} is the set of classes. We call this a classification problem.
   2. Alternatively, we may set $Y = \mathbb{R}$ and try to infer the weight of the fruit. This is a regression problem.

   b) **Example for a loss that can be used for a classification task**:

   $$L_(\bar{y_i}, y_i) = \begin{cases} 0 & \text{if } \bar{y_i} = y_i \\ 1 & \text{otherwise} \end{cases}$$

   It assumes that all types of mistakes are equally severe. Aother option is for example the **Hinge loss**:

   $$L_(\bar{y_i}, y_i) = \max(0, 1 - \bar{y_i} * y_i) \ ,$$

   where $Y = \{-1, 1\}$.
   Example for a loss that can be used regression:

   $$L_(\bar{y_i}, y_i) = \sum_{i=1}^{n} (\bar{y_i} - y_i)^2$$

   Here large deviations are more severe mistakes than small ones.

2. **Underfitting and Overfitting**

   ▷ The model in **Figure 1.1** is over-fitting. While the model performs well on he training data it does not generalize well to the new inputs in the test set, as can be seen by the classification error on the test set starting to increase while the classification error om the training set further decreases.

   ▷ The model in **Figure1.2** is under-fitting. The classification error on the training data as well as on the test data does not get very low.

   ▷ The model in **Figure1.3** models the data well and generalizes to unseen data. The classification error gets low on both sers.

3. **Independence of Random Variables**

   a) **Dices**:

   For both of the dices the probability for each of the six number is 1/6, i.e. $P(X = 6) = 1/6$ and $P(X = 3) = 1/6$. For the result of two dices there are 36 possibilities (i.e. possible results are $\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)$ $(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),$ $(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),$ $(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$), therefore $P(X = 6, Y = 3) = \frac{1}{36}$. It holds

$$p(X = 6).p(Y = 3) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = P(X = 6, Y = 3)$$

   showing that X and Y are independent.

   $P(Z = 9) = 4/36$ since there are only 4 out of 36 results that lead to a sum of 9 (i.e. $\{(3,6),(4,5),(5,4),(6,3)\}$). And $P(X = 5, Z = 9) = 1/36$ since it is just fulfilled by one of the 36 results i.e. by $(5,4)$. Thus

$$p(X = 5)p(Z = 9) = \frac{1}{6} \cdot \frac{4}{36} \neq \frac{1}{36} = P(X = 5, Z = 9)$$

   showing that Z and X are not independent.

   ==========

   Cards:

   With replacement:
   There is the same number of black and red cards in the set. Therefore $P(X = r) = P(X = b) = 1/2$ and $P(Y = r) = P(Y = b) = 1/2$. There are four possible outcomes for the colors of two cards drawn with replacement, the state space is $\{(r,b),(b,r),(r,r),(b,b)\}$. Thus, $P(X = r, Y = r) = 1/4$ and

$$p(X = r) \cdot p(Y = r) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = P(X = r, Y = r)$$

   showing that X and Y are independent.

==========

Without replacement:

For the first card we draw the number of black and red cards in the set is still the same, and thus $P(X = r) = P(X = b) = 1/2$. After the draw of a red card the numbers of red and black cards changed however resulting in $P(Y = r|X = r) = 15/31$ and

$$P(Y = r, X = r) = P(Y = r|X = r) \cdot P(X = r) = \frac{15}{31} \cdot \frac{1}{2}$$
$$\neq \frac{1}{2} \cdot \frac{1}{2} = P(X = r) \cdot P(Y = r)$$

showing that X and Y are not independent in this case.

b) The joint density is given by

$$P(x, y) = P_x(x) \cdot P_y(y) = \begin{cases} \frac{x^2}{21}.y, & \text{if } 1 \leq x \leq 4 \text{ and } 0.5 \leq y \leq 1.5 \\ 0, & \text{otherwise} \end{cases}$$

c) The marginal probability distribution of X is given by

$$P_x(X = 1) = P_{x,y}(X = 1, Y = 0) + P_{x,y}(X = 1, Y = 1) = \frac{3}{8} + \frac{1}{8} = \frac{1}{2}$$

and

$$P_x(X = 0) = P_{x,y}(X = 0, Y = 0) + P_{x,y}(X = 0, Y = 1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2} \ .$$

Analogously we get

$$P_y(Y = 1) = P_y(Y = 0) = \frac{1}{2} \ .$$

But for the joint probability mass function we have e/g.

$$P_{x,y}(X = 0, Y = 0) = \frac{1}{8} \neq \frac{1}{2} \cdot \frac{1}{2} = P_x(X = 0) \cdot P_y(Y = 0)$$

Therefor, X and Y are not independent.

4. **Statistical properties of KGs and statistical relational learning (SRL) tasks**

a) **Block structure** refers to the property where entities can be divided into distinct groups (blocks), such that all the members of a group have similar relationships to members of other groups. For example, we can group some

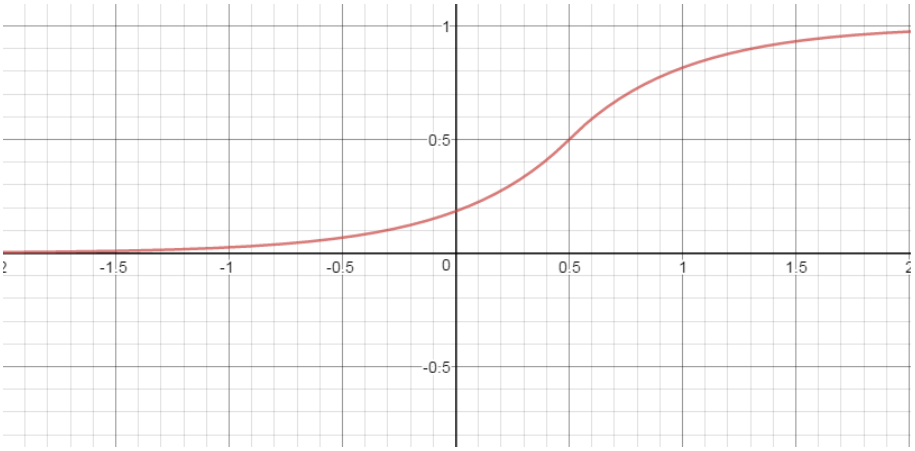
actors, such as Leonard Nimoy and Alec Guinness, into a science fiction actor block, and some movies, such as Star Trek and Star Wars, into a science fiction movie block, since there is a high density of links from the scifi actor block to the scifi movie block.
Homophily corresponds to tendency of entities to be related to other entities with similar characteristics. e.g., US-born actors are more likely to star in US-made movies.
Long-rage statistical dependencies can span over chains of triples and involve different types of relations. e.g, the citizenship of a person depends statistically on the city where he/she was born and involves a path over multiple entities (Leonard Nimoy, Boston, USA) and relations (bornIn, locatedIn, citizenOf).

b) Typical SRL tasks are **link prediction, entity resolution** and **link-based clustering**.

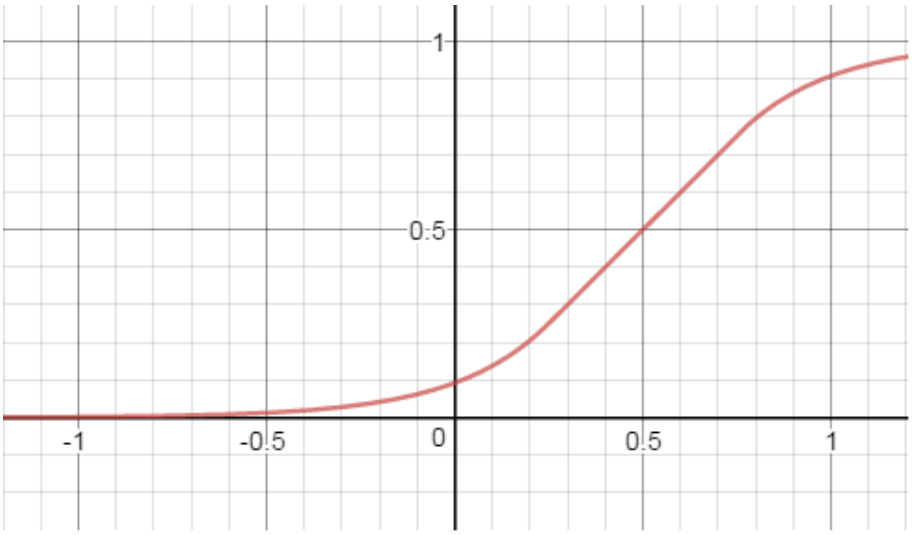
5. **Score-based and Probabilistic models for KGA**

a) Let f be a score function where $f(e_i, r_k, e_j)$ represents the models confidence that the triple (e_i, r_k, e_j) exists. **W**hile **Probabilistic models** use f to directly model the probability of the existence of a certain triple, **score-based models** optimize f under different criteria, e.g. by maximizing the margin between existing and non-existing triples.

b) A plot of $sig_{\frac{1}{2}}$ is given by:



A plot of $sig_{\frac{1}{4}}$ is given by:



The good value for ϵ can be estimated based on **cross-validation**.