Chapter 1

Linear Regression

1.1 Fitting a robust regression model using RANSAC

Algorithm 1: RANdom SAmple Consensus (RANSAC) algorithm

```
1 begin
2     repeat
3         Select a random number of examples to be inliers and fit the model;
4         Test all other data points against the fitted model and add those points
            that fall within a user-given tolerance to the inliers;
5         Refit the model using all inliers;
6         Estimate the error of the fitted model versus the inliers;
7     until the performance meets a certain user-defined threshold or if a fixed
        number of iterations was reached;
8 end
```

Chapter 2

Clustering

The goal of clustering is to find a natural grouping in data so that items in the same cluster are more similar to each other than to those from different clusters.

2.1 Prototype-based clustering

Prototype-based clustering means that each cluster is represented by a prototype, which is usually either the **centroid** (average) of similar points with continuous features, or the **medoid** (the most representative or the point that minimizes the distance to all other points that belong to a particular cluster) in the case of categorical features.

2.1.1 k-means clustering

Algorithm 2: The k-means algorithm

begin
  Randomly pick k centroids from the examples as initial cluster centers;
  repeat
    Assign each example to the nearest centroid, $\mu^{(i)}, j \in \{1, \ldots, k\}$;
    Move the centroids to the center of the examples that were assigned to it;
  until *the cluster assignments do not change or a user-defined tolerance or maximum number of iterations is reached*;
end

2.1.2 k-means++

Algorithm 3: The k-means++ algorithm

begin
  Initialize an empty set, M, to store the k centroids being selected;
  Randomly choose the first centroid from the input examples and $M \leftarrow \mu^{(j)}$;
  repeat
    For each example, $\mathbf{x}^{(i)}$, that is not in M, find the minimum squared distance, $d(x^{(i)}, M)^2$, to any of the centroids in M;
    To randomly select the next centroid, $\mu^{(p)}$, use a weighted probability distribution equal to
  until *k centroids are chosen*;
end

Chapter 3

Others

3.1 Distance

3.1.1 Euclidean distance