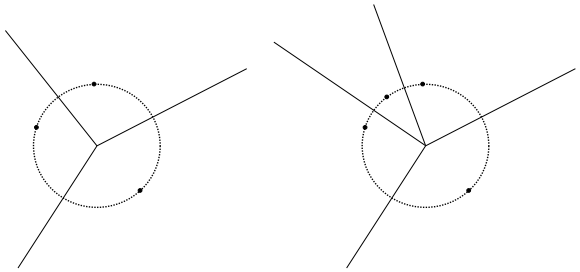
Group 6: Timm Behner, Philipp Bruckschen, Patrick Kaster, Markus Schwalb
MA-INF 4111 - Intelligent Learning and Analysis Systems: Machine Learning
Exercise Sheet 4

1. Voronoi diagrams

(a) Proof by induction over the number of points on a circle with radius r in $\mathbb{R}^2$:

Induction start: $n = 3$ Three point lying on a circle with radius r are non-collinear and the resulting cells are unbounded (figure 1a).

Induction step: $n \to n+1$ Let the induction assumption be true for n points. Let p, q be two abitrary neighboring points on the circle. We insert a new point o exactly at the intersection of the bisector of p, q and the circles radius r. The resulting cells are unbounded (figure 1b).



(a) induction start (b) induction step

Figure 1: proof by induction over points on a circle

(b) Let $\mathrm{Vor}(P)$ denote the Voronoi diagram of points P. Euler's formula for planer graphs states that

$$v - e + f = 2\,,$$

where v, e, f denote the number of vertices, edges and faces of the graph.

The number of faces is $f = |P| = n = \#\text{cells}\,(\mathrm{Vor}(P))$. To make use of Euler's formula, we have to construct a planar Graph from $\mathrm{Vor}(P)$, by adding an additional vertex to the diagram, situated at infinity, and connecting every infinite edge (from the unbounded cells) to this vertex (figure 2).

We note that each edge in $\mathrm{Vor}(P)$ has exactly two incident vertices and each vertex of $\mathrm{Vor}(P) + \infty$ has at least degree $(\deg)$ 3. Therefore we conclude

$$\sum_{\text{vertex} \in \mathrm{Vor}(P)+\infty} \deg(\text{vertex}) = 2e \geq 3(v+1) \tag{1}$$

Since we inserted our additional point ∞, Euler's Formula for our constructed graph is:

$$(v+1) - e + f = (v+1) - e + n = 2$$

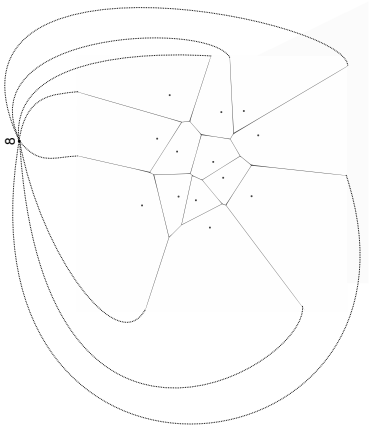


Figure 2: augmenting Voronoi diagram to planar graph

Inserting this into (1) gives us:

$$2e \geq 3(2+e-n) = 6+3e-3n \Rightarrow$$
$$e \leq 3n-6$$

Inserting this result into (1):

$$v \leq 2n-5$$

■

2. k-NN

(a) We considered the source of the dataset: `https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice`. Here it is stated that the categorical always represent some kind of scale and correspond to integer values on that scale. So we change the values given in the input file from letters to integer numbers and the boolean values "true" and "false" to 1 and 0. Now we can apply the eucledian distance function:

$$\delta(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

(b) See `output_knn.log` and `README.md`.

(c) See `output_leave_one_out.log` and `README.md`.

■

3. Hypothesis Testing

(a) $n = 100, \#'' + '' = 87$

$$\begin{aligned}
\text{error}_S(h) &= \frac{n - \#'' + ''}{n} = \frac{13}{100} = 0.13 \\
\text{error}_{\text{true}}^{95\%}(h) &= \text{error}_S(h) \pm z_N \sqrt{\frac{\text{error}_S(h)(1 - \text{error}_S(h))}{n}} \\
&= 0.13 \pm 1.96 \sqrt{\frac{0.13(1 - 0.13)}{100}} \approx [0.064, 0.196]
\end{aligned}$$

(b) With a probability of approximately 95% the true error of the hypothesis h lies within the above interval.

■

5. AUC

Use probability based ranking as described in the lecture. Assume that for every sample $x \in S$ there is a teacher output classifying x to be a positive or negative sample. Define H^+ to be the half-space that contains the most positive samples and H^- to be the other half-space. If both sides have an equal number of positive samples then choose one side to be H^+ at random. Now the probability that a sample $x \in S$ is positive is $p^+(x) = \frac{{}^+n_S^{H^i}}{n_S^{H^i}}$ where $i \in \{+, -\}$, i is the half-space where sample x is located in, ${}^+n_S^{H^i}$ is the number of all samples in H^i that are ranked positive and $n_S^{H^i}$ is the number of all samples in H^i. Obviously this would only lead to 2 ranks and would be quite a poor ranking.
A better ranking would be to consider the distance of x to the hyperplane H. The further away a point x in H^+ is away from H, the better it is ranked (one hopes that the further away a point is from the border the more sure one can be that is in the correct half-space). On the other side a point x in H^- is ranked higher the closer it is to H (here one hopes that a misclassified sample will be close to the border rather than being far away on the wrong side). Consider the distances between points x and the hyperplane H to be between x and a point on the hyperplane that yields the smallest distance in the euclidean distance measure. In general this will lead to more ranks than 2. (unless all points are located equally far away from either side of H). ■