Analyse et Traitement de l'Information

TP2: Probabilities and Statistics.
High-dimensional Data.

1 Probabilities and Statistics

1. For the table of joint probability, calculate the next values:

X	Y	$p_{XY}(x, y)$
0	0	$\frac{1}{4}$
0	1	$\frac{1}{2}$
1	0	$\frac{1}{8}$
1	1	$\frac{1}{8}$

   - $p_X(x)$
   - $p_Y(y)$
   - $p_{X|Y}(x|y = 0)$
   - $p_{Y|X}(y|x = 1)$

2. There are given two Gaussian distributions: $\mathcal{N}(15, 81)$ and $\mathcal{N}(36, 144)$. For each distribution generate 10 000 samples and plot the corresponding histograms. Show schematically at the histograms:

   (a) expected value;

   (b) variance;

   (c) standard deviation;

   (d) explain each parameter as you understand it and give its mathematical formula;

   (e) explain the difference between the histograms.

2 High-dimensional Gaussian Distribution

Generate $n = 10,000$ samples from a $\mathfrak{D}$-dimensional Gaussian distribution centered at $\mathbf{0}$ with covariance I, the identity matrix. Then, compute the 2-norm of each sample $\boldsymbol{x} = (x_1, \ldots, x_{\mathfrak{D}})^T$ by $\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{\mathfrak{D}} x_i^2}$.

- For each $\mathfrak{D} \in \{1, 10, 100\}$, plot the histogram of $\|\boldsymbol{x}\|$.
- Comment on the effect of $\mathfrak{D}$ on the distribution of $\|\boldsymbol{x}\|$.

3 "Hubness" of High Dimensional Data

Generate 100 samples in $\mathfrak{D}$-dimensional space from the zero-mean Gaussian distribution with its variance given by $I_{\mathfrak{D}}$, the $\mathfrak{D} \times \mathfrak{D}$ identity matrix. Compute the k-NN of each sample with $k = 5$. Count the number of occurrence N_i of each sample i in the k-NN of all the other samples. N_i can also be understood as the degree of the node i in the k-NN graph.

1. Plot $max(N_i)$ over $\mathfrak{D} = 1, \ldots, 100$.
2. Plot N_i over i for $\mathfrak{D} = 100$.
3. If a sample occurs frequently among the k-NN of the dataset, it can be referred to as a "hub". Explain the effect of $\mathfrak{D}$ on the occurrence of such hubs.

4 Distribution of Pair-wise Distances

Generate $n = 1000$ $\mathfrak{D}$-dimensional samples uniformly from the hyper-cube $[0, 1]^{\mathfrak{D}}$. Compute the pair-wise distances from each sample to all the other samples.

- For each $\mathfrak{D} \in \{1, 10, 100\}$, plot the *histogram* of the pair-wise distances. Explain the effect of $\mathfrak{D}$ on the distribution of the pair-wise distances.
- For each $\mathfrak{D} \in \{1, 5, 10, 50, 100\}$, compute the average distance $d_{NN}(\mathfrak{D})$ from a random sample to its nearest neighbour (NN). Plot $d_{NN}(\mathfrak{D})$ as a function of $\mathfrak{D}$. In a high dimensional space (e.g., $\mathfrak{D} = 100$), do you think that the nearest neighbour of a point $\mathbf{x}$ is still *local*?

Submission

Please archive your report and codes in "Prénom Nom.zip" (replace "Prénom" and "Nom" with your real name), and upload to "`Upload TP2 - Probabilities and Statistics. High-dimensional Data`" on `https://moodle.unige.ch` before Monday, October 14 2019, 23:59 PM. Note, the assessment is mainly based on your report, which should include your answers to all questions and the explanations of your experimental results.

Supplements

1. Define and explain what is a norm, a distance, a k-NN, and a Voronoi diagram.

2. Define and explain what is a random variable, a probability, a distribution, a cumulative distribution function, an expected value, the variance.
3. Present some distributions, their properties and applications.
4. Present the inverse theorem and its applications.
5. Present the curse of dimensionality.