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	$p_{X,Y}(x,y)$
0	0	1/2
0	1	1/8
1	0	1/4
1	1	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: $\mathcal{N}(15, 64)$ and $\mathcal{N}(36, 121)$ (CAREFULL, the second paramter is the VARIANCE not the standard deviation). 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 Simulations by using acceptance-rejection method

Given X a random variable with a continuous density function $f(\cdot)$, when we use the inversion theorem to sample the distribution of X (slides 31-33 of ATI.02), finding an explicit formula for $F^{-1}(y)$ is not always possible. A way to do is to find another density function $g(\cdot)$ easier to sample, and "close enough" from $f(\cdot)$, such that the ratio $f(x)/g(x)$ is bounded:

A constant $c > 0$ exists, such that $\frac{f(x)}{g(x)} < c$

The acceptance-rejection algorithm is as follows:

1. Sample a positive number y, from the distribution of the random variable Y of density $g(\cdot)$,
2. Sample a number u belonging to [0; 1], from the uniform distribution of the random variable U, independent from Y ,
3. If $u \leq \frac{f(y)}{cg(y)}$, accept y as a sample x for the random variable X, else, reject the sample y and go back to 1.

Such that in the end we have $P(X = y) = P(Y = y|U \leq \frac{f(Y)}{cg(Y)})$.

(optional) Questions - part 1

1. Show that $P(U \leq \frac{f(Y)}{cg(Y)}|Y = y) = \frac{f(y)}{cg(y)}$.
2. Show that $P(U \leq \frac{f(Y)}{cg(Y)}) = 1/c$.
3. Show that $P(Y = y|U \leq \frac{f(Y)}{cg(Y)}) = F(y)$. Which proves that the acceptance-rejection method is a correct sampler for X (we can also show that is is a good one, in terms of small amount of rejected samples, when c is close to 1).

Application to $\mathcal{N}(0, 1)$:

The density function of $X \sim \mathcal{N}(0, 1)$ is $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ and we will approximate the half right part of the normal density function by the exponential function of rate 1. The two considered densities are:

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, x \geq 0$$

$$g(x) = e^y, y \geq 0.$$

And the algorithm to generate samples following the positive right part of the normal distribution is a follows:

Sampling algorithm for positive normal samples:

1. sample y with the exponential distribution using the inverse theorem ($y = -\ln(u)$ where u is uniformly sampled),
2. sample u from the uniform distribution on [0, 1] independently from y,

3. if $u \leq \frac{f(y)}{cg(y)}$, set $x = y$, else restart from 1.

Questions - part 2:

1. Show that $c = \sqrt{\frac{2e}{\pi}}$ (you have to study the function $\frac{f}{g}$).
2. Implement the algorithm.
3. Suggest an algorithm to sample all the normal distribution $\mathcal{N}(0,1)$ (including the negative values).

3 High-dimensional Gaussian Distribution

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

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

4 Two Lines

Generate 50 samples uniformly on each of the following two line segments

$$l_1 : -1 \leq x \leq 1, y = -1$$
$$l_2 : -1 \leq x \leq 1, y = 1$$

- Plot the histogram of the pairwise distances among those 100 samples (4950 pairwise distances) and explain the plot;
- Add a small high-dimensional Gaussian noise to each sample. The dimension of the noise is 100. The covariance matrix of the noise is $\sigma^2 I$, where $\sigma = 0.05$. Plot the histogram of the pairwise distances again. (If necessary, try using different dimensions of the noise.) Comment your discovery.

5 Distribution of Pair-wise Distances

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

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

6 Distribution of angles

From $R^{\mathcal{D}}$ sample W, X, Y, Z iid from distribution of mean 0 (eg $[-1, 1]^D$)

- Study the distribution of angle(X-Y, Z-W).

  Remark : use of inner prod after normalisation. Should concentrate around $\frac{\pi}{2}$.
- Study the distribution of angle(X-Y, Z-Y).

  Remark : use of inner prod after normalisation. Should concentrate around $\frac{\pi}{3}$.
- Explain your protocol, describe what you see when $\mathcal{D}$ grow large and (bonus) try providing a statistical explanation.

Note: You may also relate this with quasi-orthogonality and random projections.

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 18 2021, 23:59 PM. Note, the assessment is mainly based on your report, which should include your answers to all questions and the experimental results. *Importance is given on the mathematical explanations of your works and your codes should be commented*

Supplements

Make sure that you understand all the following terms and theorems:

1. a norm, a distance, a k-NN, and a Voronoi diagram.
2. a random variable, a probability, a distribution, a cumulative distribution function, an expected value, the variance.
3. the inverse theorem and its applications.
4. the curse of dimensionality.

Also study some well-known distributions (standard, uniform, ...), their properties and applications.