Exercises
Data Mining: Learning from Large Data Sets
FS 2016

LAS Group, Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `http://las.ethz.ch/teaching/dm-f16/`
Email questions to:
Carlos Cotrini, ccarlos@inf.ethz.ch

Series 2, Oct 14th, 2016 (Locality Sensitive Hashing)

It is not mandatory to submit solutions and sample solutions will be published after one week. If you choose to submit your solution, please send an e-mail from your `ethz.ch` address with subject `Exercise2` containing a PDF (LaTeX or scan) to `ccarlos@inf.ethz.ch` until Wednesday, Oct 19th 2016.

Problem 1 (Locality Sensitive Hashing for the Cosine Distance):

(a) In this question, you'll prove that random hyperplanes can be used for locality sensitive hashing of vectors $\mathbf{v} \in \mathbb{R}^n$ when using the cosine distance. In particular, consider the family of hash functions

$$\mathcal{H} = \{h(\mathbf{v}) = \text{sign}(\mathbf{w}^T\mathbf{v}) \text{ for some } \mathbf{w} \in \mathbb{R}^n \text{ s.t. } ||\mathbf{w}||_2 = 1\}$$

Every hash function in $\mathcal{H}$, when applied to some vector $\mathbf{v} \in \mathbb{R}^n$, computes the dot product of $\mathbf{v}$ with some unit length vector $\mathbf{w}$, and returns $+1$ if the product is positive, and -1 otherwise (we ignore the case $\mathbf{w}^T\mathbf{v} = 0$ here). The vector $\mathbf{w}$ is distributed uniformly on the unit sphere $\{\mathbf{w} \in \mathbb{R}^d \mid ||\mathbf{w}|| = 1\}$. Suppose that we pick a hash function uniformly at random from $\mathcal{H}$. Show that

$$Pr([h(\mathbf{u}) = h(\mathbf{v})]) = 1 - \text{angle}(\mathbf{u}, \mathbf{v})/\pi$$

where $\text{angle}(\mathbf{u}, \mathbf{v}) \in [0, \pi]$ denotes the angle (in radians) between two n-dimensional vectors.

(b) Write a Python script that will calculate the probability of two vectors having the same hash as a function of their cosine distance. More precisely, you should sample

- N pairs of random vectors in $\mathbb{R}^d$. You can for example use the uniform distribution on $[-1, 1]^d$, or a normal $\mathcal{N}(\mathbf{0}, I_{d\times d})$.
- M random hash functions in $\mathbb{R}^d$. Instead of using the family in the previous part, you can use sketches (vectors sampled uniformly from $\{-1, +1\}^d$), or you could try normally distributed vectors.

Then, for each of the N pairs you should compute the fraction of the M hash functions that agree on them. You are free to pick N and M and d as you wish. Create scatter plots that show your results. Do they relate to the result from the previous section? What happens when you change the dimension of the data?

Solution 1(a):

Fix any non-zero vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and consider the plane on which they lie. We will use the two following facts

- The probability that the projection of $\mathbf{w}$ lies in some circular sector is proportional to the angle of that sector (all angles are equally probable).
- For any $\mathbf{z}$ we have that $\mathbf{z}^T\mathbf{x} = \bar{\mathbf{z}}^T\mathbf{x}$ and $\mathbf{z}^T\mathbf{y} = \bar{\mathbf{z}}^T\mathbf{y}$, where $\bar{\mathbf{z}}$ is the projection of $\mathbf{z}$ on this plane.

We will restrict our attention to the case when the angle between $\mathbf{x}$ and $\mathbf{y}$ is at most 90 degrees, because $P_{\mathbf{w}}(h_{\mathbf{w}}(\mathbf{x}) = h_{\mathbf{w}}(\mathbf{y})) = P_{\mathbf{w}}(h_{\mathbf{w}}(\mathbf{x}) \neq h_{\mathbf{w}}(-\mathbf{y}))$ and $\text{angle}(\mathbf{x}, -\mathbf{y}) = 180 - \text{angle}(\mathbf{x}, \mathbf{y})$. We will compute the probability $P_{\mathbf{w}}(h_{\mathbf{w}}(\mathbf{x}) = h_{\mathbf{w}}(y) = 1)$. Then, because $(-\mathbf{w})^T\mathbf{x} = -\mathbf{w}^T\mathbf{x}$ by symmetry we get that $P_{\mathbf{w}}(h_{\mathbf{w}}(\mathbf{x}) = h_{\mathbf{w}}(y)) = 2P_{\mathbf{w}}(h_{\mathbf{w}}(\mathbf{x}) = h_{\mathbf{w}}(y) = 1)$. Consider the set

$$H = \{\mathbf{z} \in \mathbb{R}^n \mid \mathbf{z}^T \frac{1}{2}(\mathbf{x} + \mathbf{y}) \geq 0\}$$

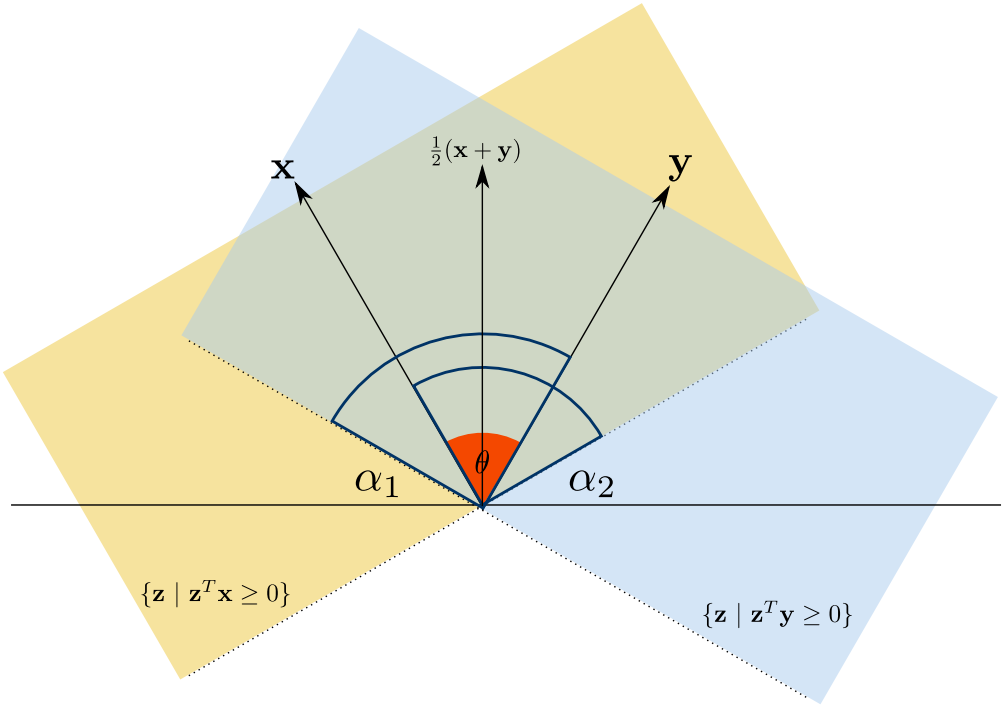

Figure 1: The plane on which $\mathbf{x}$ and $\mathbf{y}$ lie. H is the set above the horizontal line. The blue angles without interior are 90 degrees.

All $\mathbf{w}$ which satisfy $\mathbf{w}^T\mathbf{x} \geq 0$ and $\mathbf{w}^T\mathbf{y} \geq 0$ belong to H (sum up the two inequalities). Let us denote by $R(\alpha)$ the circular sector corresponding to the angle α. Using the notation on Figure 1 we have that

$$P_{\mathbf{w}}(h_{\mathbf{w}}(\mathbf{x}) = h_{\mathbf{w}}(y) = 1) = \underbrace{P(\mathbf{w} \in H)}_{1/2} - P_{\mathbf{w}}(\mathbf{w} \in R(\alpha_1)) - P_{\mathbf{w}}(\mathbf{w} \in R(\alpha_2))$$

Because the probability of falling into a sector is proportional to the angle of that sector

$$\begin{aligned} P_{\mathbf{w}}(\mathbf{w} \in R(\alpha_1)) + P_{\mathbf{w}}(\mathbf{w} \in R(\alpha_2)) &= 1/2 - \frac{1}{360}(90 + 90 - \theta) \\ &= \theta/360 \end{aligned}$$

And this completes the proof because

$$\begin{aligned} P_{\mathbf{w}}(h_{\mathbf{w}}(\mathbf{x}) = h_{\mathbf{w}}(y)) &= 2P_{\mathbf{w}}(h_{\mathbf{w}}(\mathbf{x}) = h_{\mathbf{w}}(y) = 1) \\ &= 2(1/2 - \theta/360) \\ &= 1 - \theta/180 \end{aligned}$$

Solution 1(b):

The implementation is provided in `lsh-cosine.py`. We show below some graphs generated with normally distributed hashes and uniformly distributed data. The other options should give similar results (sketches in very low dimensions won't work very nicely as there are only a few possible hashes). We indeed see the behaviour proven in the previous section. Moreover, as we increase the dimensions we see that the angle between two random samples (under the distributions considered) becomes more concentrated.