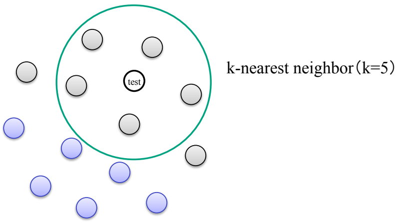
近傍法における距離・類似度尺度のデータ中心化 ーハブネスの軽減ー

鈴木郁美　　統計的機械学習センター　特任研究員

Research Highlight

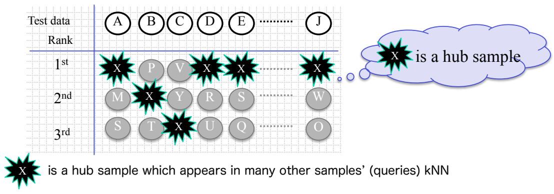
- Similarity measures based on inner products are popular measures in NLP and other machine learning tasks.
- kNN does not work well for high-dimensional data.
- [Radovanovic et al. 2010] pointed out that **hubs** emerge in high-dimensional space.
  A Hub is a sample which is similar to many other samples in a dataset.
  The presence of hubs can deteriorate the accuracy of kNN-based classification.
- [Radovanovic et al. 2010] showed that samples close to / similar to the data centroid tend to become hubs.
- **We show that simple "Data Centering" technique can reduce hubs** and improve kNN based classification performance.

Classification based on kNN



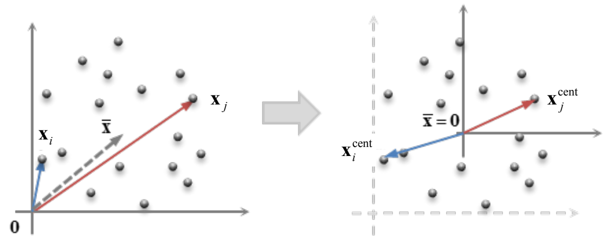
A label of a test sample is predicted by labels of k training samples which are most similar to the test sample.

What is a hub?



X is a hub sample which appears in many other samples' (queries) kNN

Data Centering $\mathbf{x}^{\text{cent}} = \mathbf{x} - \bar{\mathbf{x}}$



Similarity between the i and j-th samples is measured by inner product of their feature vectors.

Before Centering: $\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle$

After Centering: $\langle \mathbf{x}^{\text{cent}\,(i)}, \mathbf{x}^{\text{cent}\,(j)} \rangle = \langle \mathbf{x}^{(i)} - \bar{\mathbf{x}}, \mathbf{x}^{(j)} - \bar{\mathbf{x}} \rangle$

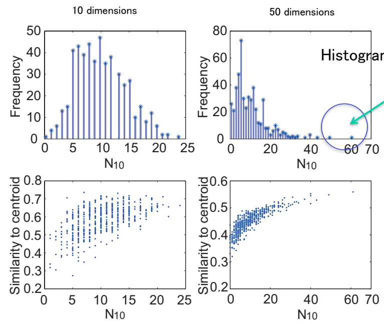
Emergence of Hubs

A sample which is similar to the data centroid tends to become a hub

Synthetic dataset
- 500 samples with 10 and 50 dimensions
- Cosine similarity is used to measure similarity between samples

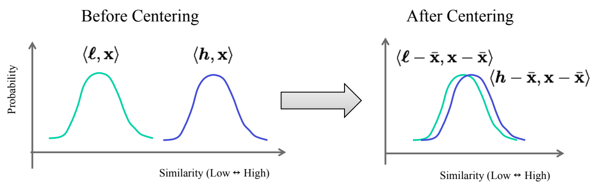
Evaluate N_{10} value for each sample in a dataset
- N_k is the number of times a sample appears in other samples' kNN

N_k Value is large for hub samples



N_{10}: Number of times a sample appears in other samples' 10NN

Why Data Centering Reduces Hubs?



① Select two points $\boldsymbol{h}$ and $\boldsymbol{\ell}$, such that $\langle \boldsymbol{h}, \boldsymbol{\mu} \rangle > \langle \boldsymbol{\ell}, \boldsymbol{\mu} \rangle$

Then, compare similarity (inner product) for the selected samples ($\boldsymbol{h}$ and ℓ) with other samples $\mathbf{x}$.

② Before Centering : How different is the mean of the distribution $\langle \boldsymbol{h}, \mathbf{x} \rangle$ and $\langle \boldsymbol{\ell}, \mathbf{x} \rangle$?

$$\mathrm{E}[\langle \boldsymbol{h}, \mathbf{x} \rangle] - \mathrm{E}[\langle \boldsymbol{\ell}, \mathbf{x} \rangle] = \langle \boldsymbol{h}, \mathrm{E}[\mathbf{x}] \rangle - \langle \boldsymbol{\ell}, \mathrm{E}[\mathbf{x}] \rangle = \langle \boldsymbol{h}, \boldsymbol{\mu} \rangle - \langle \boldsymbol{\ell}, \boldsymbol{\mu} \rangle > 0$$

∴ The mean of two distributions are different : $\mathrm{E}[\langle \boldsymbol{h}, \mathbf{x} \rangle] > \mathrm{E}[\langle \boldsymbol{\ell}, \mathbf{x} \rangle]$

③ After Centering : What become of the mean difference of the distribution $\langle \boldsymbol{h} - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}} \rangle$ and $\langle \ell - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}} \rangle$?

$$\langle \boldsymbol{h}^{\text{cent}}, \mathbf{x}^{\text{cent}} \rangle = \langle \boldsymbol{h} - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}} \rangle = \langle \boldsymbol{h}, \mathbf{x} \rangle - \langle \boldsymbol{h}, \bar{\mathbf{x}} \rangle - \langle \mathbf{x}, \bar{\mathbf{x}} \rangle + \|\bar{\mathbf{x}}\|^2 \qquad \langle \ell^{\text{cent}}, \mathbf{x}^{\text{cent}} \rangle = \langle \ell - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}} \rangle = \langle \ell, \mathbf{x} \rangle - \langle \ell, \bar{\mathbf{x}} \rangle - \langle \mathbf{x}, \bar{\mathbf{x}} \rangle + \|\bar{\mathbf{x}}\|^2$$

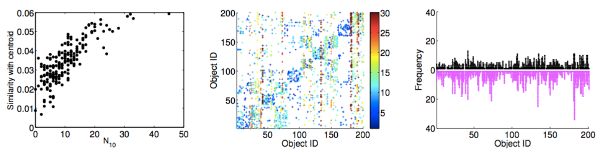
$$\mathrm{E}[\langle \boldsymbol{h} - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}} \rangle] - \mathrm{E}[\langle \ell - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}} \rangle] = \mathrm{E}[\langle \boldsymbol{h}, \mathbf{x} \rangle] - \mathrm{E}[\langle \boldsymbol{h}, \bar{\mathbf{x}} \rangle] - \mathrm{E}[\langle \ell, \mathbf{x} \rangle] + \mathrm{E}[\langle \ell, \bar{\mathbf{x}} \rangle] = 0$$

∴ The mean of two distributions are not different : $\mathrm{E}[\langle \boldsymbol{h} - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}} \rangle] = \mathrm{E}[\langle \ell - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}} \rangle]$

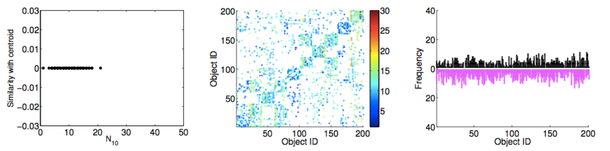
Centroid vector: $\bar{\mathbf{x}} = \sum_{i=1}^{N} \mathbf{x}_i$

Experiments

Multi-cluster Analysis with Reuters Transcribed



(a) Before centering: N_{10} vs. inner product similarity to the data centroid
(b) Before centering: kNN matrix
(c) Before centering: Breakdown of N_{10} by class match/mismatch between objects and neighbors



(d) After centering: N_{10} vs. inner product similarity to the data centroid
(e) After centering: kNN matrix
(f) After centering: Breakdown of N_{10} by class match/mismatch between objects and neighbors

Reuters Transcribed data. (a), (d): scatter plot of the N_{10} value of objects and their similarity to centroid. (b), (e): kNN matrices. The points are colored according to the N_{10} value of object x; warmer colors indicate higher N_{10} values. (c), (f): the number of times (y-axis) an object (whose ID is on the x-axis) appears in the 10 nearest neighbors of objects of the same cluster (black bars), and those of different clusters (magenta).