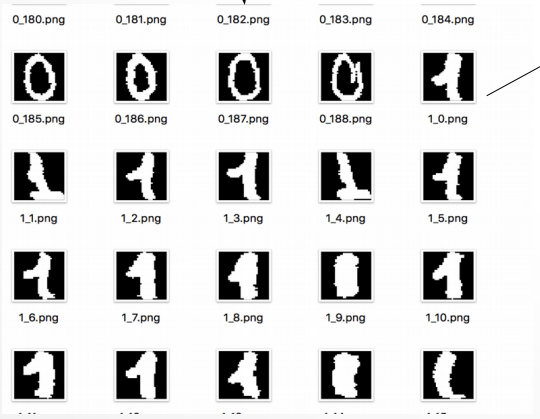
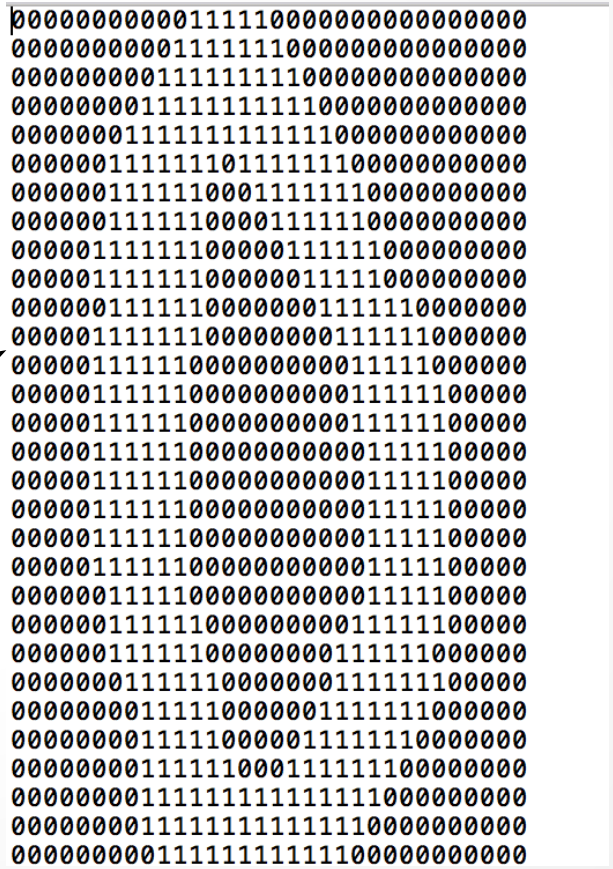
Chiffre Reconnaissance





Chiffre Reconnaissance

- Gagner des données (kaggle.com)
- Analyser des données et préprocession (scaling et PCA)
- Générer modèle(KNN et SVM)
- Evaluation

Dataset

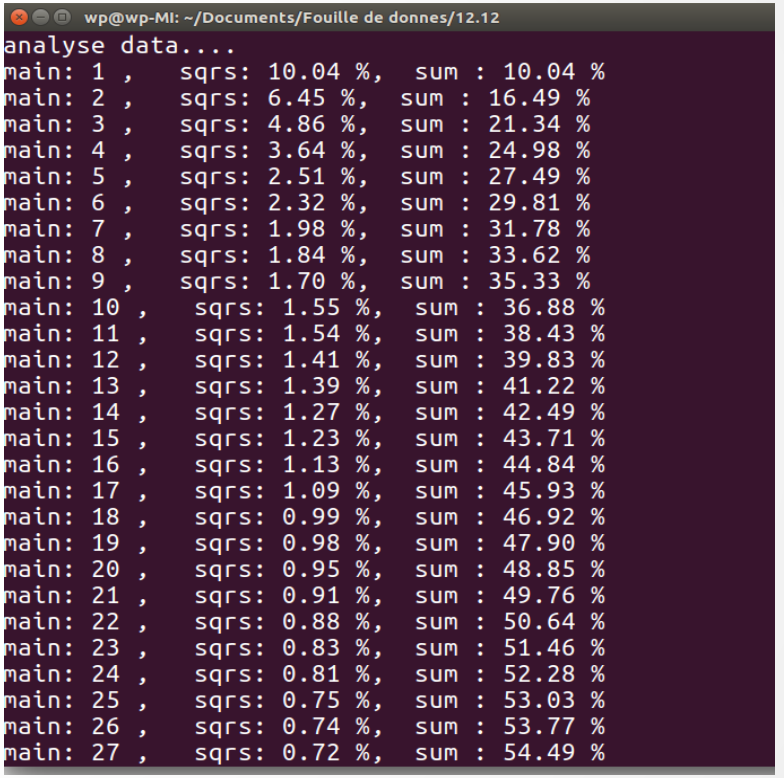
- Source : Kaggle.com
- 70MB+ dataset, parmi lesquels on a extrait 42000 échantillons, et 785 feutures par échantillon.
- Each image is 28 pixels in height and 28 pixels in width, for a total of 28*28=784 pixels in total. Each pixel-value is an integer between 0 and 255.

Pre-processing

- Normalisation de données(scaling)
- X = $\frac{X - \mu}{\sigma}$
- Réduction de dimension
- PCA (Principal Component Analysis)

PCA(Analyse en composantes principales)

- Analyser des données
- Calcule le variance pour chaque dimension
- Mise en ordre (max→ min)
- Faire l‘addition jusqu‘au somme de variances représente plus 95%
- Sauvegarder les premiers 320 dimensions , Couper les restes dimensions



Principe de PCA

- Le tableau X (M lignes N columns)
- Covariance matrix $C = \frac{1}{m} X X^T$
- Eigenvector et eigenvalue par SVD (singular value decomposition)

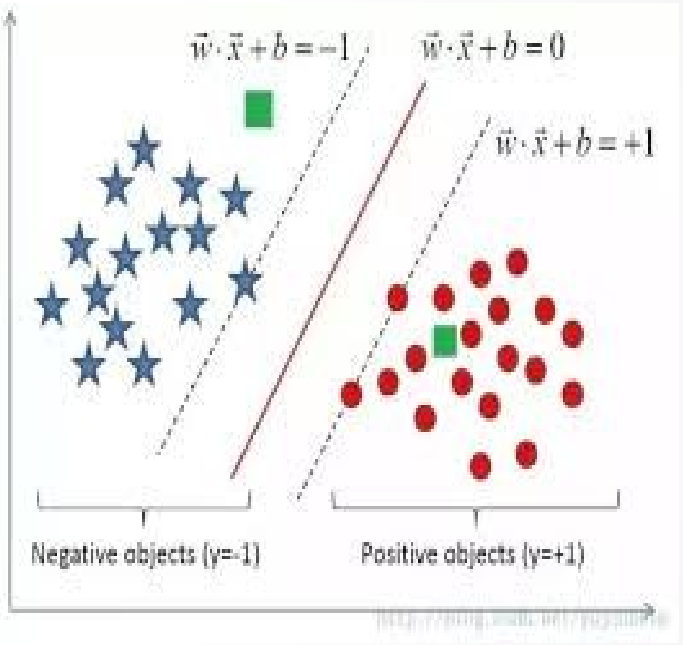
$$A = UDV^T,$$

- Mise en ordre pour eigenvalue, sauvegarder les premiers k lignes comme un nouveux matrix P.
- Y = PX est le resultat , et qui a eu k dimensions

$$\begin{aligned} D &= \frac{1}{m} Y^T Y \\ &= \frac{1}{m} (PX)^T (PX) \\ &= \frac{1}{m} X^T P^T P X \\ &= X^T (\frac{1}{m} P^T P) X \\ &= X^T C X \end{aligned}$$

SVM (Support Vector Machine)

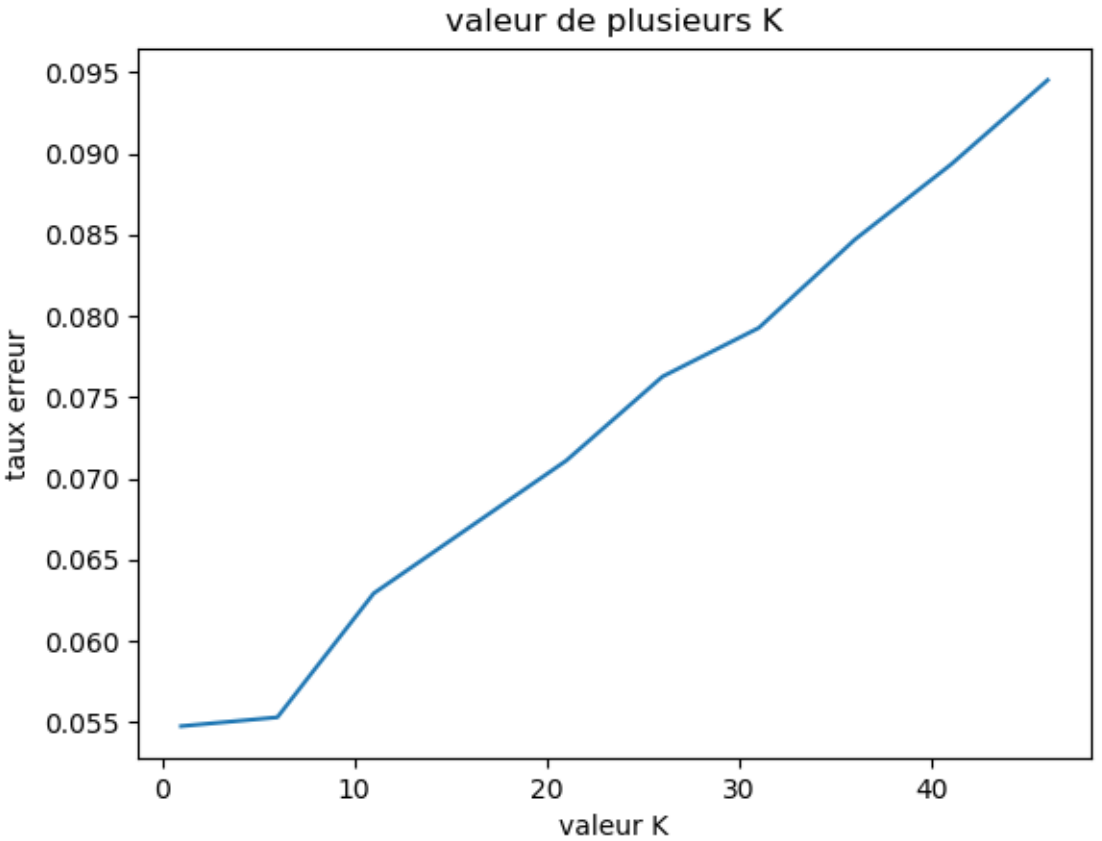
- Kernel : Linear
- Linear separability
- Haute-dimensions
- Le tableau est clairsemé
- Économic le temps , et obtenir de bons résultats

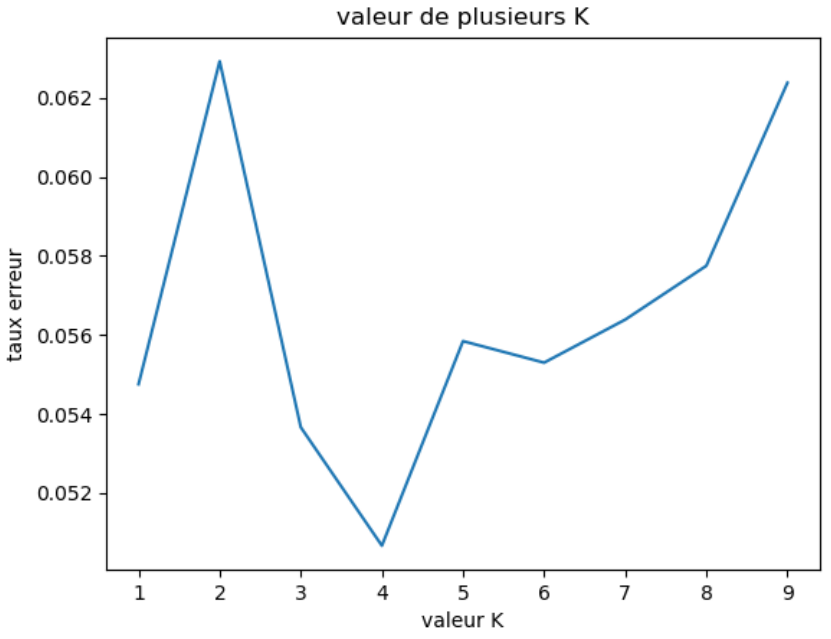


Entrainement / Validation

- 70% pour l’entrainement
- 30% pour les tests

K nearest neighbors




valeur de plusieurs K
taux erreur
0.062
0.060
0.058
0.056
0.054
0.052
1
2
3
4
5
6
7
8
9
valeur K

Précision de modèle

- KNN : 0.9493326069190956
- SVM : 0.9588667937891583

Retropection

- Séparation de Entrainement / Validation
  - Mal séparer les 2 ensembles avec la taille espéré
  - On aurait du utiliser train_test_split de sklearn

- Pour KNN
  - Manque validation croisé
  - Améliorer la fonction de distance
    - Avec pondération peut-etre