

Rapport de projet de reconnaissance d'images

Lucas David & Théo Legras

1 Implémentation et utilisation du classificateur à distance minimum (DMIN)

On choisi d'implémenter ce classificateur sous forme d'une classe, ceci étant le plus courant et le plus pratique pour encapsuler les comportements et stocker les données nécessaires (extrait du fichier `dmin.py`) :

```
1 import numpy as np
2
3 class DMIN:
4     def __init__(self):
5         self.data = np.array([])
6         self.label = np.array([])
7         self.n_samples = 0
8
9     def fit(self, data, label):
10        self.data = data
11        self.label = label
12        self.n_samples = len(set(label))
13        return self
14
15    def predict(self, data):
16        return [self.label[np.argmin(np.sum(np.subtract(self.data, data[iterator])
17        ** 2, axis=1))] for iterator in range(0, len(data))]
18
19    def score(self, data, label):
20        return np.count_nonzero(self.predict(data) == label) / len(data)
```

L'utilisation se résumera à l'instanciation de DMIN à l'appel de `DMIN.fit` et selon l'usage l'appel de `DMIN.predict` et `DMIN.score`. En particulier, on peut donc déterminer le taux de réussite via la fonction membre `DMIN.score(<données à tester>, <labels correspondants>)`.

Dans le cas de nos données de développement, on obtient un score de 68,80% pour une exécution de 96,45 secondes. Il est toujours intéressant de noter que si on teste l'ensemble d'entraînement, on obtient le score parfait... On verra plus tard que ce n'est pas le cas de tous les algorithmes car cela peut être un indicateur d'*overfitting* (du surapprentissage ou de la surinterprétation), c'est-à-dire correspond trop étroitement aux données.

2 Utilisation de l'analyse en composantes principales (PCA) et application à DMIN

L'utilisation de l'Implémentation de la PCA (`sklearn.decomposition.PCA`) est plutôt simple. On peut choisir via le paramètre `n_components` le nombre de dimensions à garder, si $0 \leq n_components < 1$, on indique la proportion des données à garder en variance (%).

Nous nous choisissons de faire nos tests en modulant en variance plus qu'en nombre de dimensions car la notion de variance peut être mise en parallèle avec la perte de précision à postériori de la PCA. De plus cela permet d'écarter les cas de réductions dans des nombres dimensions proches (e.g. passer de 760 dims. à 700 dims. ne signifie pas grand chose alors que de 100 dims. à 40 dims. à un impact visible). Effectivement, selon le modèle on n'obtiendra pas la même courbe "Variance des données par rapport aux nombres de dimensions" et ce n'est généralement pas linéaire.

Globalement, nous appliquons dans les grandes lignes la PCA et DMIN sur les données réduites comme suit :

```
1 import numpy as np
2 from sklearn.decomposition import PCA
3
4 X = np.load('data/trn_img.npy')
5 Y = np.load('data/trn_lbl.npy')
6 devX = np.load('data/dev_img.npy')
7 devY = np.load('data/dev_lbl.npy')
8
9 pca = PCA(n_components=0.5) # Ici on garde 50% de la variance des données.
10 reducedX = pca.fit_transform(X) # On colle au modèle et on transforme X.
11 print('Dimensions: {}'.format(reducedX)) # On peut afficher le nombre de
    dimension en valeur.
12 reducedDevX = pca.transform(reducedDevX) # On transforme devX.
13
14 dmin = DMIN() # On utilise DMIN.
15 dmin.fit(reducedX, Y)
16 print('Score: {}'.format(dmin.score(reducedDevX, devY)))
```