Machine Learning Programming Exercise 5: Regularized Linear Regression and Bias vs Variance

adapted to Python language from Coursera/Andrew Ng

25 septembre 2020

1 Objectif

Nous allons dans ce TP estimer la diminution de la surface de la mer d'Aral (partagée entre le Kazakhstan au nord et l'Ouzbékistan au sud), à partir des couvertures satellites de 1973 et de 1987 (données USGS).

L'objectif est d'écrire des scripts permettant de mettre en œuvre un système basé sur différentes approches de machine learning. Ce système visera l'estimation sur chaque image de la surface de la mer d'Aral et en déduira sa diminution au cours du temps. Ces scripts devront ainsi suivre la chaîne générale décrite en cours (à l'exception de la phase de captation) :

— prétraitements
— extraction des descripteurs
— apprentissage d'un modèle de classement
— classement des pixels

Pour commencer avec cette séance, vous aurez besoin de télécharger le « stater code »disponible sur le lien Moodle du cours.

Dans cet exercice, il vous est demandé de fournir un rapport regroupant les réponses aux questions et vos analyses en plus de vos codes.

2 Fichiers inclus dans le starter code pour cette séance

— **Aral1973_Clean.jpg** - fichier image de 1973 de la mer d'Aral ;
— **Aral1987_Clean.jpg** - fichier image de 1987 de la mer d'Aral ;
— **loadImages.py** - charge les images de la mer d'Aral ;
— **displayFeatures2d.py** - permet d'afficher les valeurs des descripteurs en 2d ;
— **displayFeatures3d.py** - permet d'afficher les valeurs des descripteurs en 3d ;
— **displayImageLabel.py** - permet l'affichage de l'image des labels à partir d'un vecteur de labels prédits ;
— **selectFeatureVectors.py** - mise en forme des descripteurs
⋆ **aralsea_main.py** - script Python qui vous servira de point de départ pour cette séance ;
⋆ **preprocessing.py** - fonction qui charge les deux images, les prétraitent, les mets en forme pour la suite,
⋆ **unsupervisedTraining.py** - apprentissage d'un classifieur non supervisé selon deux modèles ;
⋆ **unsupervisedClassifying.py** - classement d'un classifieur non supervisé selon deux modèles.

[⋆] indique les fichiers que vous devez completer.

3 Examen des données, prétraitements et extraction des descripteurs

Question 1: Compléter la fonction **preprocessing()** permettant de charger les deux images et de les afficher. indication : **loadImages()**.

Question 2: Que peut-on déduire de l'analyse de ces images ? Comment sont codées les valeurs des pixels ?

Question 3: Les images sont en couleurs et nous allons utiliser les composantes des couleurs comme descripteurs de classification. Quelle est la dimension de l'espace des descripteurs lorsque l'image est codée en RGB (RVB) ?

Question 4: Tronquer de la même manière les deux images pour faire disparaître le texte dans la partie haute et basse des images.

4 Constitution d'un ensemble de données d'apprentissage

Question 1: Compléter la fonction **preprocessing()** par une partie qui vise à constituer une base d'apprentissage. Pour cela, réalisez un sous-échantillonnage de l'image de 1973 avec un pas de 500 points (indication : **selectFeatureVectors.py**). Quel est le nombre de données d'apprentissage ?

5 Première analyse des données

Question 1: Dans **preprocessing**, visualiser en 2D et 3D les valeurs des descripteurs. La visualisation en 2D des vecteurs d'apprentissage se fait par la fonction **displayFeatures2d(feat)** et en 3D **displayFeatures3d(feat)**.

Question 2: Décrivez explicitement ces graphiques en expliquant ce qu'ils représentent, en donnant leurs caractéristiques pour les histogrammes et les nuages de points (différents groupes ? à quelles informations peut-on les relier ?)

6 approche non supervisée par la méthode des k-means

Pour cette première approche non-supervisée, nous allons mettre en œuvre un système basé sur l'algorithme des k-moyennes (k-means), variantes de la méthode des centres mobiles. Créer un script reprenant ce que vous avez fait dans la section précédente.

Question 1: Décrire avec vos mots la méthode des centres mobiles.

Question 2: Appeler et compléter le script **unsupervisedTraining.py** qui permet réaliser un apprentissage non supervisé du modèle de classement (classifieur) à l'aide de la fonction **KMeans()** du package scikit-learn (puissant package de Machine Learning pour python). Comment utilise-t-on cette fonction ? Quels en sont les paramètres de contrôle importants ?

Question 3: Compléter **aralsea_main.py** pour prédire les labels sur la base d'apprentissage grâce au modèle appris.

Question 4: Il est maintenant possible de visualiser les valeurs des descripteurs d'apprentissage et leur appartenance à l'une des classes (un des « clusters »). Nous utiliserons la fonction : **displayFeatures2d** et **displayFeatures3d**. Compléter **aralsea_main.py**. Faire varier le paramétrage de la fonction KMeans et analyser les différences et la qualité de l'apprentissage.

Question 5: A ce stade, vous avez choisi les hyper-paramètres de vos méthodes de machine learning et obtenu le modèle de classement, nous pouvons alors utiliser les données d'apprentissage pour classifier l'ensemble des deux images. Il faut calculer les descripteurs sur toute l'image (mise sous la forme d'une matrice) puis utiliser le modèle de classement issu de l'apprentissage du k-means). Compléter les scripts **aralsea_main.py** et **unsupervisedClassifying.py**.

Question 6: Pour visualiser le résultat de classification, il faut que chaque classe prédite puisse être affectée au pixel correspondant de l'image. Cette image est calculée et visualisée par la fonction **displayImageLabel.py**. Compléter **aralsea_main.py**

Question 7: Après avoir identifiée sur l'image la couleur de la classe (et son numéro) de la zone correspondant à la mer d'Aral, il est possible d'en estimer la surface (i.e. le nombre de pixels) sur les deux images et d'en estimer l'évolution. Quelle est approximativement cette évolution en % ? Pour répondre à cette question compléter **aralsea_main.py**

Question 8: Analyser en fonction des paramètres du kmeans, la création de classe dans l'espace des descripteurs et donner l'impact que cela a sur le pourcentage estimé ?

7 Approche non-supervisée par GMM

Question 1: Décrire en détails, l'apprentissage du classifieur non supervisé basé sur les mélanges de gaussiennes.

Question 2: Compléter votre code pour apprendre un classifieur basé des mélanges de gaussiennes

7.1 Some questions, you have to answer...