td2a_enonce_cl_reg_anomaly

August 12, 2022

1 2A.data - Classification, régression, anomalies - énoncé

Le jeu de données Wine Quality Data Set contient 5000 vins décrits par leurs caractéristiques chimiques et évalués par un expert. Peut-on s'approcher de l'expert à l'aide d'un modèle de machine learning.

[1]:
```
%matplotlib inline
import matplotlib.pyplot as plt
```

[2]:
```
from jyquickhelper import add_notebook_menu
add_notebook_menu()
```

[2]:
```
<IPython.core.display.HTML object>
```

1.1 Les données

On peut les récupérer sur github...data_2a.

[3]:
```
from ensae_teaching_cs.data import wines_quality
from pandas import read_csv
df = read_csv(wines_quality(local=True, filename=True))
df.head()
```

[3]:

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality	color
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	red
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	red
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	red
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	red
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	red

1.2 Exercice 1 : afficher la distribution des notes

La fonction hist est simple, efficice.

1.3 Exercice 2 : séparation train / test

La fonction est tellement utilisée que vous la trouverez rapidement.

1.4 Exercice 3 : la variable couleur n'est pas numérique

M... OneHotEncoder.

[4]:

1.5 Exercice 3 : premier classifieur

Vous trouverez aussi tout seul. Quelques fonctions pourront vous aider à évaluer le modèle confusion_matrix, classification_report.
Beaucoup mieux.

1.6 Exercice 4 : courbe ROC

Quelques aides...

[5]:
```
from sklearn.metrics import roc_curve, auc

# labels = pipe.steps[1][1].classes_
# y_score = pipe.predict_proba(X_test)

fpr = dict()
tpr = dict()
roc_auc = dict()
# for i, cl in enumerate(labels):
#     fpr[cl], tpr[cl], _ = roc_curve(y_test == cl, y_score[:, i])
#     roc_auc[cl] = auc(fpr[cl], tpr[cl])
```

[6]:
```
# fig, ax = plt.subplots(1, 1, figsize=(8,4))
# for k in roc_auc:
#     ax.plot(fpr[k], tpr[k], label="c%d = %1.2f" % (k, roc_auc[k]))
# ax.legend();
```

1.7 Exercice 5 : anomalies

Une anomalie est un point aberrant. Cela revient à dire que sa probabilité qu'un tel événement se reproduise est faible. Un modèle assez connu est EllipticEnvelope. On suppose que si le modèle détecte une anomalie, un modèle de prédiction aura plus de mal à prédire. On réutilise le pipeline précédent en changeant juste la dernière étape.

[7]:
```
from sklearn.covariance import EllipticEnvelope
```

1.8 Exercice 6 : régression

La note est numérique, pourquoi ne pas essayer une régression.

[8]:
```
from sklearn.ensemble import RandomForestRegressor
```

1.9 Exercice 7 : intervalle de confiance

Comment constuire un intervalle de confiance avec un classifieur et un régresseur. Rien de théorique, juste des idées et un peu de bidouille.

[9]: