

Liens

[Blog \(./blog/main_oooo.html\)](#)

[Index \(./genindex.html\)](#)

[Module \(./py-modindex.html\)](#)

[\(./api/index.html\)](#)

Contenu (index.html)

[Introduction \(introduction.html\)](#)

[Quelques rappels \(rappel.html\)](#)

[Lectures sur le machine learning \(lectures/index.html\)](#)

[Devinettes et problèmes \(i_questions.html\)](#)

[Bibliographie \(biblio.html\)](#)

[API \(api/index.html\)](#)

[Exemples \(i_examples.html\)](#)

[Blog Gallery \(blog/blogindex.html\)](#)

[En diagonal \(i_index.html\)](#)

[Glossary \(glossary.html\)](#)

Information

[FAQ \(./i_faq.html\)](#)

[Glossaire \(./glossary.html\)](#)

[README \(./README.html\)](#)

[Changements \(./filechanges.html\)](#)

[Licence \(./license.html\)](#)

- « Receiver Oper... (gyexamples/plots/plot_roc.html)
- Menu dans un notebook » (notebooks/menu_notebook.html)

[Source \(_sources/all_notebooks.rst.txt\)](#)

Galleries de notebooks

[Notebooks Coverage \(all_notebooks_coverage.html#l-notebooks-coverage\)](#)

- Astuces de datascientist
- Digressions machine learnsques
- Notebooks commencés en cours
- Exemples et solutions
- Visualisation

Astuces de datascientist

Menu dans un notebook
(notebooks/menu_notebook.html#menunotebookrst)

Insérer un
menu en
javascript.

```
<|Pythc
<div id="a2">run previous cell, wait fo
</div>
function repeat_indent_string(n){
  for (i=0; i<n; i++){
    repeat_indent_string(n-1);
  }
}
```

(notebooks/menu_notebook.html)

Visualiser un arbre de décision
(notebooks/decision_tree_visualization.html#decisiontreevisualizationrst)

Les arbres
de décision
sont des
modèles
intéressants
car ils
peuvent
être
interprétés.
Encore
faut-il
pouvoir les
voir.

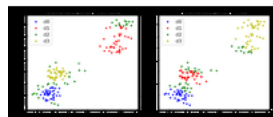
```
<jyquickhelp
<div id="M6f444d344a324e179d04d35
```

(notebooks/decision_tree_visualization.html)

Digressions machine learnsques

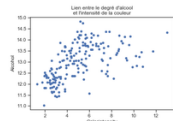
K-Means contraint
(notebooks/constraint_kmeans.html#constraintkmeansrst)

Les k-
means
construisent
des clusters
qui ne sont
pas
forcément
équilibrés.
Comment
modifier
l'algorithme
original
pour que ce
soit le cas ?



(notebooks/constraint_kmeans.html)

Notebooks commencés en cours

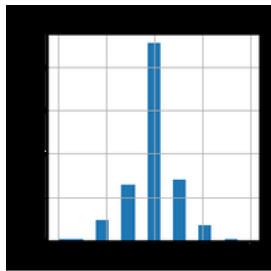


Un exemple simple de
régression linéaire
(notebooks/2019-01-
25_linreg.html#linregrst)

Un jeu de données, une régression linéaire et quelques
trucs bizarres. Le jeu de données est Wine Data Set .

(notebooks/2019-
01-
25_linreg.html)

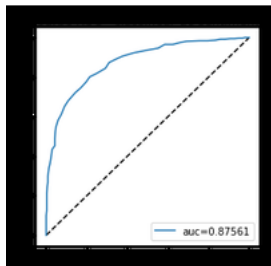
Exemples et solutions



(notebooks/wines_knn_split.html)

Base d'apprentissage et de test
(notebooks/wines_knn_split.html#winesknnsplitrst)

Le modèle est estimé sur une base d'apprentissage et évalué sur une base de test.



(notebooks/text_sentiment_wordvec.html)

Classification de phrases avec word2vec
(notebooks/text_sentiment_wordvec.html#textsentsimentwordvecrst)

Le texte est toujours délicat à traiter. La langue est importante et plus le vocabulaire est étendu, plus il faut de données. Le problème qui suit est classique, on cherche à catégoriser des phrases en sentiment positif ou négatif. Ce pourrait être aussi classer des spams. Le problème le plus simple : une phrase, un label.

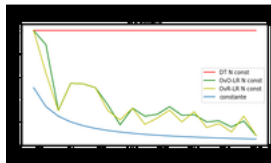
Classification multi-classe (notebooks/wines_multi.html#winesmultirst)

On cherche à prédire la note d'un vin avec un classifieur multi-classe.

47

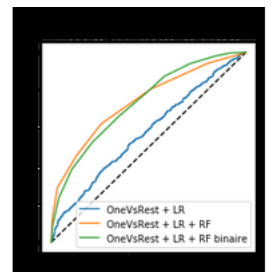
(notebooks/wines_multi.html)

Classification multi-classe et jeu mal équilibré
(notebooks/artificiel_multiclass.html#artificielmulticlassrst)



(notebooks/artificiel_multiclass.html)

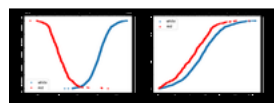
Plus il y a de classes, plus la classification est difficile car le nombre d'exemples par classe diminue. Voyons cela plus en détail sur des jeux artificiels produits par make_blobs.



(notebooks/wines_multi_stacking.html)

Classification multi-classe et stacking
(notebooks/wines_multi_stacking.html#winesmultistackingrst)

On cherche à prédire la note d'un vin avec un classifieur multi-classe puis à améliorer le score obtenu avec une méthode dite de stacking.



Classifications et courbes ROC
(notebooks/wines_color_roc.html#winescolorrocrst)

La courbe ROC est une façon de visualiser la performance d'un classifieur ou plutôt sa pertinence. Voyez comment.

(notebooks/wines_color_roc.html)

Sondage @OdoxaSondages Décryptage
J'ss sure c'est Raouf qui a cree ce # 🤖

Des mots aux sacs de mots
(notebooks/artificiel_tokenize_features.html#artificieltokenizefeaturesrst)

La tokenisation consiste à découper un texte en *token*, l'approche *sac de mots* consiste à compter les occurrences de chaque mot dans chaque document de la base de données.

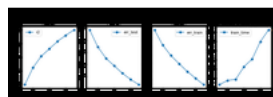
(notebooks/artificiel_tokenize_features.html)

-0 1

Découpage stratifié apprentissage / test
(notebooks/wines_knn_split_strat.html#winesknnsplitstratrst)

Lorsqu'une classe est sous-représentée, il y a peu de chances que la répartition apprentissage test conserve la distribution des classes.

(notebooks/wines_knn_split_strat.html)



Factorisation de matrice et recommandations
(notebooks/movielens_fm.html#movielensfmrst)

Le notebook utilise la factorisation de matrice pour calculer des recommandations sur la base movielens. On utilise le jeu de données ml-latest-small.zip.

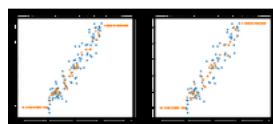
(notebooks/movielens_fm.html)

```
col_int  0  1  2  3  4  5
cat_text
cat31
dataframe tbody tr th:only
vertical-align: middle;
dataframe tbody tr th:only
vertical-align: middle;
```

Hashing et catégories
(notebooks/artificiel_category_hash.html#artificielcategoryhashrst)

Le hashing est utilisé lorsque le nombre de catégories est trop grand.

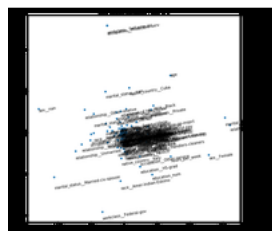
(notebooks/artificiel_category_hash.html)



Intuition derrière l'érosion des modèles
(notebooks/artificiel_shape.html#artificielshaperst)

La plupart des modèles de machine learning doivent être rafraîchi régulièrement. Quelques intuitions derrière ce phénomène.

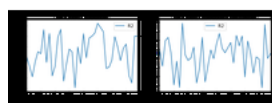
(notebooks/artificiel_shape.html)



(notebooks/adult_cat.html)

Jeu de données avec des catégories (notebooks/adult_cat.html#adultcatrst)

Le jeu de données Adult Data Set ne contient presque que des catégories. Ce notebook explore différents moyens de les traiter.



(notebooks/artificiel_normalisation.html)

Normalisation
(notebooks/artificiel_normalisation.html#artificielnormalisationrst)

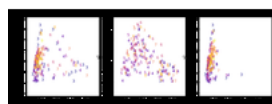
La normalisation des données est souvent inutile d'un point de vue mathématique. C'est une autre histoire d'un point de vue numérique où le fait d'avoir des données qui se ressemblent améliore la convergence des algorithmes et la précision des calculs. Voyons cela sur quelques exemples.



(notebooks/wines_knn.html)

Plus proches voisins (notebooks/wines_knn.html#winesknnrst)

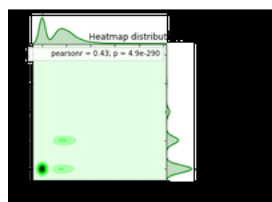
On cherche à prédire la note d'un vin avec un modèle des plus proches voisins.



(notebooks/wines_knn_acp.html)

Plus proches voisins - projection
(notebooks/wines_knn_acp.html#winesknnacprst)

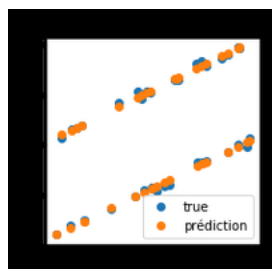
On projette le jeu de données initiale selon les premiers axes d'une analyse en composantes principales (ACP).



(notebooks/wines_knn_eval.html)

Plus proches voisins - évaluation
(notebooks/wines_knn_eval.html#winesknnevalrst)

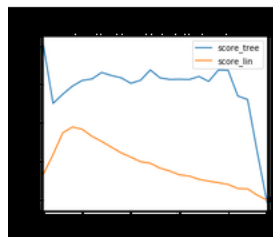
Comment évaluer la pertinence d'un modèle des plus proches voisins.



(notebooks/wines_color_linear.html)

Plusieurs modèles, données disjointes
(notebooks/wines_color_linear.html#winescolorlinearrst)

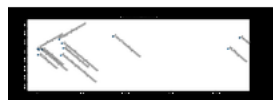
On cherche à prédire la note d'un vin mais on suppose que cette qualité est dépendante de la couleur et qu'il faudrait appliquer des modèles différents selon la couleur.



(notebooks/artificiel_duration_prediction.html)

Prédiction d'une durée
(notebooks/artificiel_duration_prediction.html#artificieldurationpredictionrst)

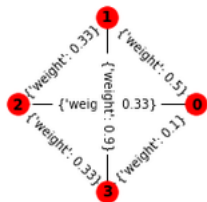
Ce notebook cherche à prédire la durée de stockage de paquets préparés par un magasin. Chaque paquet met plus ou moins de temps à être préparé. Si la commande arrive le soir et ne peut être finie avant la fin de la journée, elle est reportée sur la journée suivante. C'est la particularité de ce jeu de données



(notebooks/wines_reg.html)

Prédiction de la note des vins (notebooks/wines_reg.html#winesregrst)

Le notebook compare plusieurs de modèles de régression.



(notebooks/tinygraph_rwr.html)

Random Walk with Restart (système de recommandations)
(notebooks/tinygraph_rwr.html#tinygraphrwrrst)

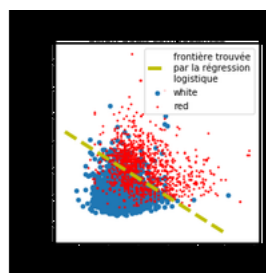
Si la méthode de factorisation de matrices est la méthode la plus connue pour faire de recommandations, ce n'est pas la seule. L'algorithme Random Walk with Restart s'appuie sur l'exploration locale des noeuds d'un graphe et produit de résultats plus faciles à interpréter.*.

Random rank
Our model: 0

(notebooks/search_rank.html)

Ranking et search engine (notebooks/search_rank.html#searchrankrst)

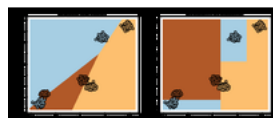
C'est un petit exemple de ranking avec un très petit jeu de données, trop petit pour que le modèle soit performant, mais le code peut être réutilisé pour des exemples de taille raisonnable. C'est à dire probablement pas pour apprendre un moteur de recherche.



(notebooks/wines_color_line.html)

Régression logistique en 2D
(notebooks/wines_color_line.html#winescolorlinerst)

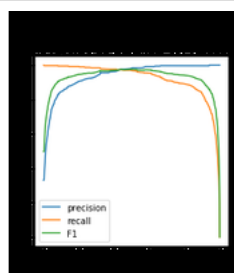
Prédire la couleur d'un vin à partir de ses composants.



(notebooks/logreg_kmeans.html)

Régression logistique et convexité
(notebooks/logreg_kmeans.html#logregkmeansrst)

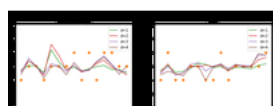
La régression logistique est un modèle de classification binaire simple à interpréter mais limité dans la gamme des problèmes qu'il peut résoudre. Limité comment ?



(notebooks/wines_color.html)

Régression logistique et courbe ROC
(notebooks/wines_color.html#winescolorrst)

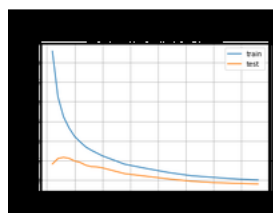
Prédire la couleur d'un vin à partir de ses composants et visualiser la performance avec une courbe ROC.



(notebooks/wines_reg_poly.html)

Régression polynômiale et pipeline
(notebooks/wines_reg_poly.html#winesregpolrst)

Le notebook compare plusieurs de modèles de régression polynômiale.



(notebooks/wines_knn_hyper.html)

Sélection des hyper-paramètres
(notebooks/wines_knn_hyper.html#winesknnhyperrst)

Le modèle des plus proches voisins KNeighborsRegressor est paramétrable. Le nombre de voisins et variables, la prédiction peut dépendre du plus proche voisins ou de k plus proches voisins. Comment choisir k ?

Tokenisation (notebooks/artificiel_tokenize.html#artificieltokenizerst)

```
thinc 7.0.4
preshed preshed-2.0.1.dist-
cymem cymem-2.0.2.dist-
murmurhash murmurhash
blac 0.9.0
spacy 2.1.4
```

La tokenisation consiste à découper un texte en *token*, le plus souvent des mots. Le notebook utilise un extrait d'un article du monde.

(notebooks/artificiel_tokenize.html)

 Traitement amélioré des catégories
 (notebooks/artificiel_category_2.html#artificielcategory2rst)

Ce notebook présente des encodage différents de ceux implémentés dans scikit-learn.

n 9c

(notebooks/artificiel_category_2.html)

 Traitement des catégories
 (notebooks/artificiel_category.html#artificielcategoryrst)

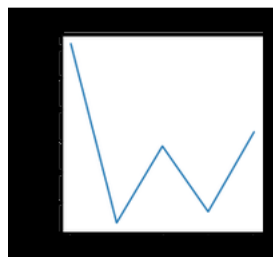
Ce notebook présente différentes options pour gérer les catégories au format entier ou texte.

n 9c

(notebooks/artificiel_category.html)


 Validation croisée (cross-validation)
 (notebooks/wines_knn_cross_val.html#winesknncrossvalrst)

Il est acquis qu'un modèle doit être évalué sur une base de test différente de celle utilisée pour l'apprentissage. Mais la performance est peut-être juste l'effet d'une aubaine et d'un découpage particulièrement avantageux. Pour être sûr que le modèle est robuste, on recommence plusieurs fois. On appelle cela la validation croisée ou cross validation.



(notebooks/wines_knn_cross_val.html)

Visualisation

30/10/2019	Galleries de notebooks — papierstat	
<div data-bbox="110 184 373 220"><pre><folium <div style="width:100%;"><div style=</pre></div> <div data-bbox="110 352 410 380">(notebooks/carte_carreau.html)</div>	<div data-bbox="501 90 961 144">Données carroyées (notebooks/carte_carreau.html#cartecarreaurst)</div> <div data-bbox="1109 90 1234 434">Les données carroyées sont des données économiques agrégées sur tout le territoire français sur des carrés de 200m de côté.</div>	
<div data-bbox="110 499 371 686"></div> <div data-bbox="110 728 410 756">(notebooks/enedis_cartes.html)</div>	<div data-bbox="501 464 961 518">Tracer une carte en Python (notebooks/enedis_cartes.html#enediscartesrst)</div> <div data-bbox="1109 464 1227 663">Le notebook propose plusieurs façons de tracer une carte en Python.</div>	
<div data-bbox="110 877 373 972"><pre>Année Nom commune dataframe tbody tr th (vertical-align: middle; dataframe tbody tr th (</pre></div> <div data-bbox="110 1045 480 1073">(notebooks/enedis_cartes_bokeh.html)</div>	<div data-bbox="501 781 1088 835">Tracer une carte en Python avec bokeh (notebooks/enedis_cartes_bokeh.html#enediscartesbokehrst)</div> <div data-bbox="1109 781 1234 1155">bokeh permet de tracer une carte sur laquelle on peut zoomer, dézoomer et qui ne dépend pas d'un service extérieur comme folium.</div>	