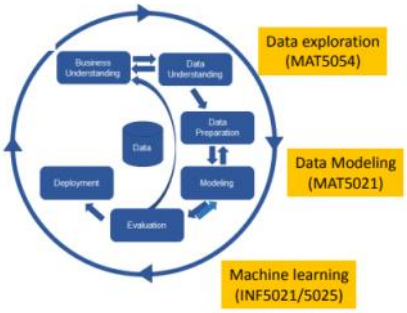
Big Picture and Good Practices.

mercredi 21 septembre 2022 14:23



Data Understanding :

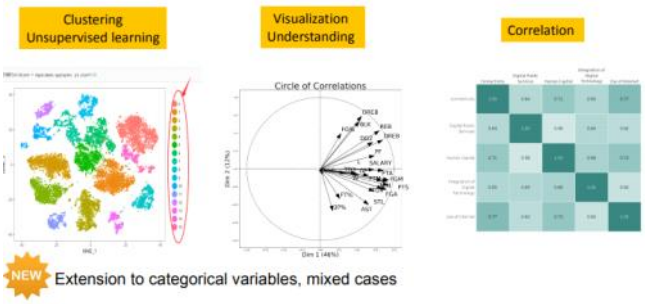
=> Data Cleansing (check data quality & handle missing values)

Check data quality:
- validity (data type/range/membership... constraints
- completeness (see below)
- consistency (same value across datasets)
- ...

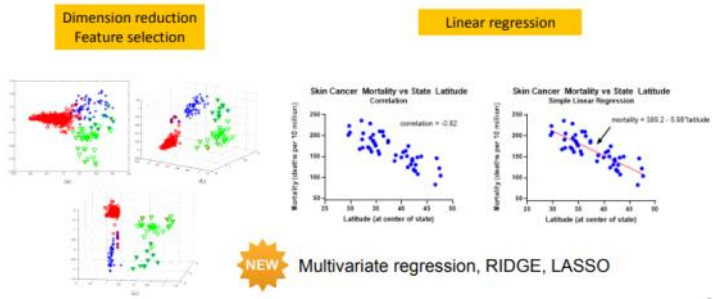
Handle missing values:
- basic: ignore data row/column
- use a constant (out of range)
- imputation: replace with mean/median
- try to find algorithms that are less sensitive to this problem

En plus de cela, on regarde le problème d'échelle et la notion de variable naturelle/supplémentaire.

=> Data Exploration (on regarde à quoi ressemblent nos données)



=> Data Modeling & Feature Engineering



Model Training

=> TOUT DE SUITE APRES LE CLEANSING

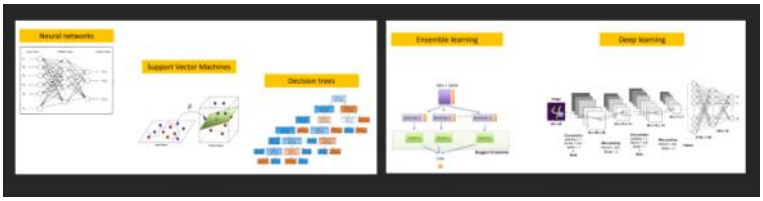
Split data into training set and test set

Depending on the algorithm, transform nominal variable into numerical ones: **one-hot-encoding**

For small or unbalanced training set: increase sample size by using **data augmentation** (e.g SMOTE)

Normalize data:
- estimate mean *m* and standard variation *s* on the training set
- for each feature/dimension : $x' = (x - m)/s$

Machine Learning **Advanced Machine Learning**



NEW /! **---- Information Fusion ----**

Performance Evaluation

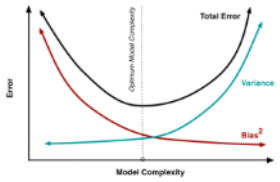
Tuning Parameters (réglages des paramètres) :

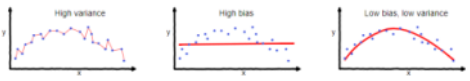
- **Pour KNN =>** K
- **Pour Decision Trees =>** max_depth, min_sample_split
- **Pour Neuronal Networks =>** number of hidden layers/cells, learning rate

Pour éviter le sur-apprentissage (et donc une mauvaise généralisation) :

- k-fold cross validation

Biais/Variance





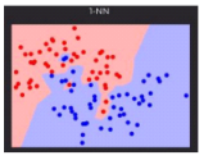
overfitting underfitting Good balance

Biais = Erreur
Variance = Collage aux données

Avec les taux d'apprentissage et la variance, on calcule la moyenne et l'écart-type.
On veut une variance faible donc un écart-type faible aussi => donc biais (taux d'erreur, donc 1 - taux de reconnaissance) élevé ? Il faut toujours un compromis !
Plus on augmente la complexité du modèle, plus le biais est faible et la variance est haute mais si la variance est trop haute, le modèle apprend trop (overfitting) donc mauvaise généralisation.

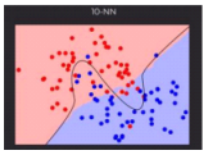
Compromis entre biais et variance :

1-NN : colle trop aux données d'entrainement => relation trop forte au training set => variance

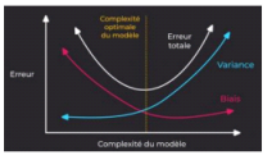


En augmentant k, on lisse nos frontière et on se détache de cette dépendance au training set => cela diminue donc la variance
Par contre en augmentant k, on crée un autre type d'erreur qui est le biais => à quel point on vise à côté de la vraie valeur d'un point considéré

10-NN : frontières de décisions très lisses, très éloignée de la zone idéale



La difficulté va être de diminuer l'erreur du modèle tout en prenant en compte le biais et la variance qui évolue de manière contraire !



- Quand on est dans le cas d'une variance excessive; réduire le nombre de dimensions réduire la variance en simplifiant la complexité
- Il faut sélectionner le bon modèle qui trouve la complexité optimal et fait ainsi la balance entre les deux.
- La manière d'entrainer le modèle est aussi important : il y a des méthodes qui minimise la prise en compte de variance non représentative du modèle.
- Méthodes d'ensemble = famille d'algorithmes qui se basent sur la combinaison de modèles à haute variance pour réduire la variance finale.

Modèle à trop grande complexité = modèle à haute variance

Overfitting et underfitting :

- La notion d'Overfitting (quelquefois traduite par surapprentissage en français) désigne le fait que le modèle que vous avez choisi est trop collé aux données d'entraînement. C'est un problème classique de data science, lorsqu'on choisit un modèle trop "flexible", c'est-à-dire avec une complexité trop élevée qui prend aussi en compte le bruit du phénomène. C'est en fait ce qui arrive aux méthodes à haute variance dont je parlais dans le chapitre précédent.
- À contrario, l'**underfitting** (ou **sous-apprentissage**) désigne une situation où le modèle n'est pas du tout assez complexe pour capturer le phénomène dans son intégralité.

Donc pour faire le lien avec le chapitre précédent, les méthodes à haute variance ont tendance à *overfitter facilement* sur du bruit non représentatif du modèle sous-jacent. À l'inverse, les algorithmes avec un biais élevé ont tendance à *underfitter plus facilement* les données d'entraînement, et donc à rater des informations importantes sur le phénomène.

Lorsqu'on veut généraliser un modèle, on a besoin qu'il n'*overfit* pas et qu'il n'*underfit* pas, qu'il soit pile entre les deux.

On considère qu'un algorithme est particulièrement puissant lorsqu'il possède cette capacité de généralisation, et qu'il peut effectuer les prédictions les plus performantes possibles avec le moins de données possibles.

Mesures de la performance :

=> taux de reconnaissance, taux d'erreur, taux de rejet
=> matrice de confusion
=> Faux négatif/faux positif

CHALLENGES :

- Bias reduction
- Explicability (X-AI)
- Increntality (continuous/lifelong learning)
- Sustainability (green AI)
- Acceptability (Human in the loop)