De la statistique classique à l'apprentissage automatique

1) De la statistique à l'apprentissage machine

Historique :

1940-70 : *Statistiques classiques*. Question associée à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.

1970s : Généralisation des premiers outils informatiques. L'analyse de données explore des données plus volumineuses.

1980s : Les systèmes experts sont supplantés par l'apprentissage automatique.

1990s : *1er changement de paradigme* : Les données ne sont plus planifiées mais sont préalablement acquises : *From Data Mining to Knowledge Discovery*.

2000s : *2eme changement de paradigme* : Le nombre de variables p explose, notamment avec les données omiques où $p>>n$. La qualité de prévision devient plus importante que la réalité du modèle devenu *boîte noire*. Problématique du fléau de la dimension.

2010s : *3eme changement de paradigme* : Le nombre d'observations n explose dans le e-commerce, la géo-localisation, … . Bases de données structurées en *cloud* et moyens de calculs regroupés en *clusters (big data)*. La notion de rapidité des algorithmes devient critique.

2020s : Explosion des usages dû à la facilité d'accès à des données massives et à des ressources calcul puissantes. → explicabilité des décisions des algorithmes pour des raisons sociétales — robustesse des décisions dans un cadre critique — embarquabilité des réseaux de neurones — cybersécurité — … les stratégies *anciennes* restent souvent intéressantes sur des données complexes (small data)

1) De la statistique à l'apprentissage machine

Historique :

1940-70 : *Statistiques classiques*. Question associée à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.

1970s : Généralisation des premiers outils informatiques. L'analyse de données explore des données plus volumineuses.

1980s : Les systèmes experts sont supplantés par l'apprentissage automatique.

1990s : *1er changement de paradigme* : Les données ne sont plus planifiées mais sont préalablement acquises : *From Data Mining to Knowledge Discovery*.

2000s : *2eme changement de paradigme* : Le nombre de variables p explose, notamment avec les données omiques où $p>>n$. La qualité de prévision devient plus importante que la réalité du modèle devenu *boîte noire*. Problématique du fléau de la dimension.

2010s : *3eme changement de paradigme* : Le nombre d'observations n explose dans le e-commerce, la géo-localisation, … . Bases de données structurées en *cloud* et moyens de calculs regroupés en *clusters (big data)*. La notion de rapidité des algorithmes devient critique.

2020s : Explosion des usages dû à la facilité d'accès à des données massives et à des ressources calcul puissantes. → explicabilité des décisions des algorithmes pour des raisons sociétales — robustesse des décisions dans un cadre critique — embarquabilité des réseaux de neurones — cybersécurité — … les stratégies *anciennes* restent souvent intéressantes sur des données complexes (small data)

1) De la statistique à l'apprentissage machine

Historique :

1940-70 : *Statistiques classiques*. Question associée à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.

1970s : Généralisation des premiers outils informatiques. L'analyse de données explore des données plus volumineuses.

1980s : Les systèmes experts sont supplantés par l'apprentissage automatique.

1990s : *1er changement de paradigme* : Les données ne sont plus planifiées mais sont préalablement acquises : *From Data Mining to Knowledge Discovery*.

2000s : *2eme changement de paradigme* : Le nombre de variables p explose, notamment avec les données omiques où $p>>n$. La qualité de prévision devient plus importante que la réalité du modèle devenu *boîte noire*. Problématique du fléau de la dimension.

2010s : *3eme changement de paradigme* : Le nombre d'observations n explose dans le e-commerce, la géo-localisation, … . Bases de données structurées en *cloud* et moyens de calculs regroupés en *clusters (big data)*. La notion de rapidité des algorithmes devient critique.

2020s : Explosion des usages dû à la facilité d'accès à des données massives et à des ressources calcul puissantes. → explicabilité des décisions des algorithmes pour des raisons sociétales — robustesse des décisions dans un cadre critique — embarquabilité des réseaux de neurones — cybersécurité — … les stratégies *anciennes* restent souvent intéressantes sur des données complexes (small data)

1) De la statistique à l'apprentissage machine

Historique :

1940-70 : *Statistiques classiques*. Question associée à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.

1970s : Généralisation des premiers outils informatiques. L'analyse de données explore des données plus volumineuses.

1980s : Les systèmes experts sont supplantés par l'apprentissage automatique.

1990s : *1er changement de paradigme* : Les données ne sont plus planifiées mais sont préalablement acquises : *From Data Mining to Knowledge Discovery*.

2000s : *2eme changement de paradigme* : Le nombre de variables p explose, notamment avec les données omiques où $p>>n$. La qualité de prévision devient plus importante que la réalité du modèle devenu *boîte noire*. Problématique du fléau de la dimension.

2010s : *3eme changement de paradigme* : Le nombre d'observations n explose dans le e-commerce, la géo-localisation, … . Bases de données structurées en *cloud* et moyens de calculs regroupés en *clusters (big data)*. La notion de rapidité des algorithmes devient critique.

2020s : Explosion des usages dû à la facilité d'accès à des données massives et à des ressources calcul puissantes. → explicabilité des décisions des algorithmes pour des raisons sociétales — robustesse des décisions dans un cadre critique — embarquabilité des réseaux de neurones — cybersécurité — … les stratégies *anciennes* restent souvent intéressantes sur des données complexes (small data)