Reconstruction des événements à l'aide de Machine Learning en physique des particules

Journée thématique "activités Machine Learning à l'IPHC"

1/10/2020

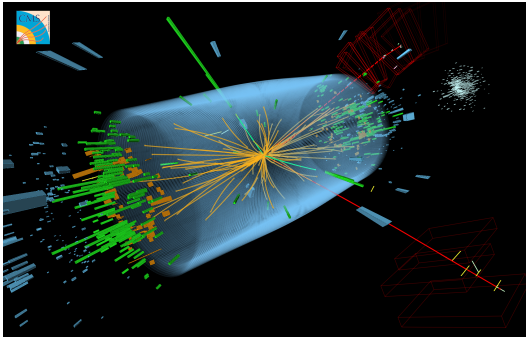
Emery Nibigira

Sommaire

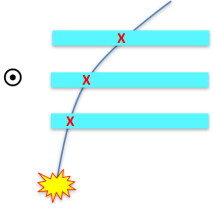
- Contexte
- Problème de classification
- Algorithmes de classification
- Applications
- Conclusion

Reconstruction des événements en physique des particules

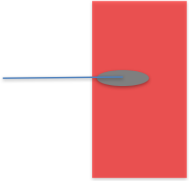
repose sur notre capacité à détecter les particules issues de chaque collision

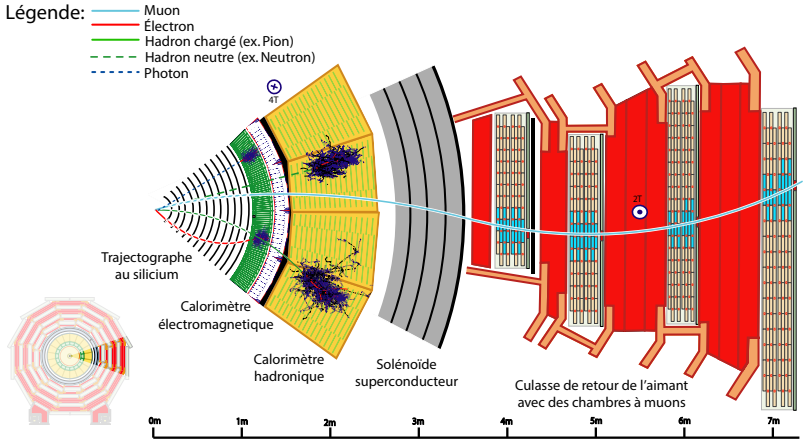


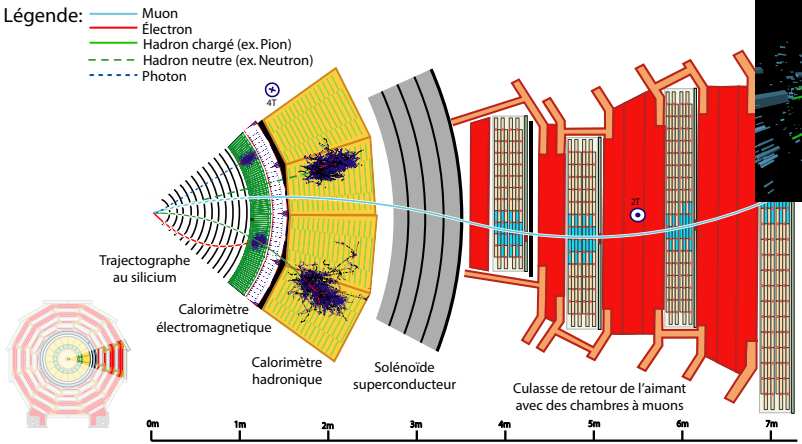
- **trajectographie**
  points de passage
  - trace
  - impulsion $p = 0.3B \times r$

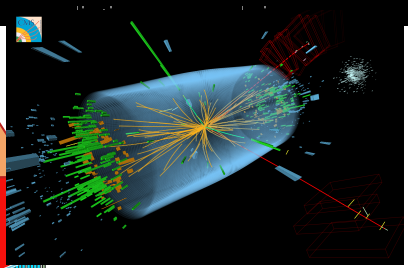


- **calorimétrie**
  dépôts d'énergie
  - énergie

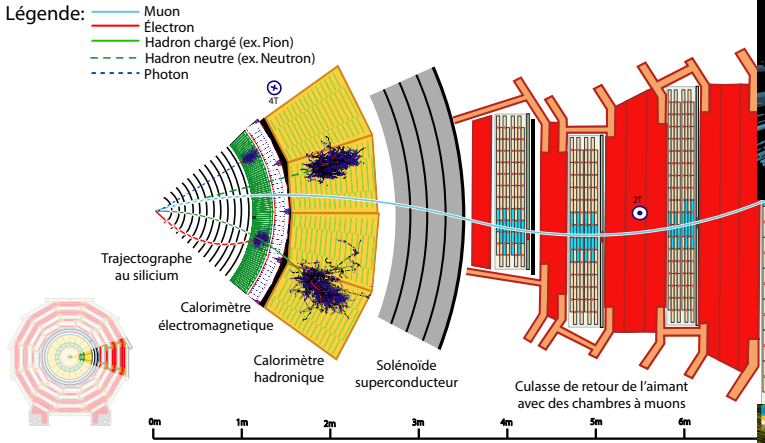


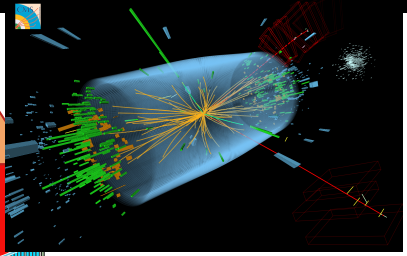

Légende:
Muon
Électron
Hadron chargé (ex. Pion)
Hadron neutre (ex. Neutron)
Photon
4T
2T
Trajectographe
au silicium
Calorimètre
électromagnetique
Calorimètre
hadronique
Solénoïde
superconducteur
Culasse de retour de l'aimant
avec des chambres à muons
0m
1m
2m
3m
4m
5m
6m
7m



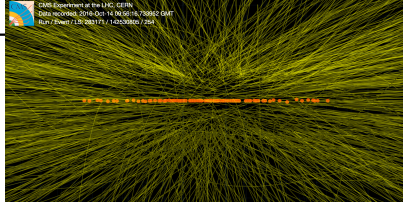


[1] Identifier les particules électrons, muons, ...





[1] Identifier les particules électrons, muons, ...



[2] Identifier le processus

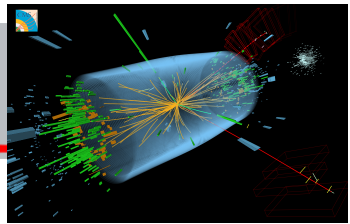
pp⟶?⟶électrons, muons, ...

Prediction: Modèle Standard, Autre modèle

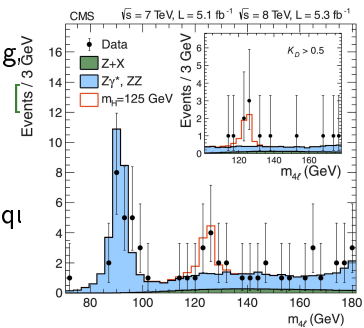
Collision pp $\Longrightarrow$ croisement entre deux **paquets de protons** (10^{11}protons chacun au LHC)

- Jusqu'à 10^9 collisions par seconde
  - ⋆ **taux de collision élevé** $\longrightarrow$ acquisition (électronique)
  - ⋆ **grande quantité de données (Big data)** $\longrightarrow$ computing, stockage
- Nécessite un système de **filtrage**:
  - on ne peut pas tout stocker
  - sélectionner que les événements intéressants pour la physique

Collision pp $\Longrightarrow$ croisement entre deux **paquets de protons** (10^{11} protons chacun au LHC)



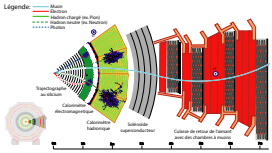
- Jusqu'à 10^9 collisions par seconde
  - ⋆ **taux de collision élevé** $\longrightarrow$ acquisition (électronique)
  - ⋆ **grande quantité de données (Big data)** $\longrightarrow$ computing
- Nécessite un système de **filtrage**:
  - on ne peut pas tout stocker
  - sélectionner que les événements intéressants pour la physiqu
- **Analyse physique** (signal vs bruit, mesure de précision)
  - données simulées et observées



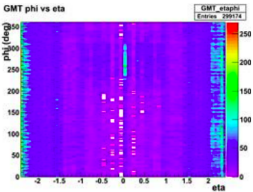
Utilisation de Machine Learning (ML)

- **Supervisé**: classification, regression
- **Non supervisé**
- ...

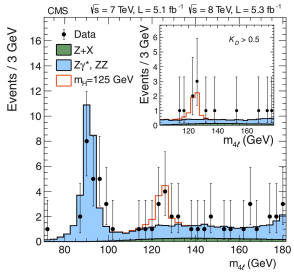
Identification, reconstruction



Monitoring



Analyse



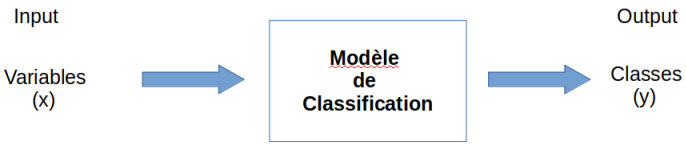
Computing

WLCG
Worldwide LHC Computing Grid

Électronique

Problème de classification

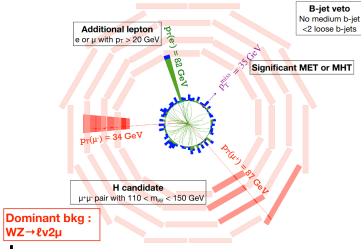
Rappel (Présentation de E. Chabert)



Prenons un exemple: Higgs

A chaque événement,
une série de variables dont on connaît les valeurs

- input: variables discriminantes (x_i)
- output: 2 classes → oui (événement Higgs) ou non



Classification: construire à un modèle, basé sur des **algorithmes visant à prédire** le groupe d'appartenance Y pour chaque valeur x_i

Problème de classification

Échantillon d'entraînement

Id	masse	impulsion	charge	classe
1	34	90	-1	Non
2	25	170	1	Non
3	84	90	1	Oui
4	50	100	-1	Non
5	22	80	-1	Oui
6	139	200	1	Non
7	75	80	1	Oui
8	43	12	-1	Non
9	40	90	1	Non
10	84	70	-1	Oui

Échantillon Test

Id	masse	impulsion	charge	classe
11	86	65	-1	?
12	115	170	1	?
13	32	11	1	?
14	45	99	-1	?
15	76	80	1	?

Principe:

1. **Entraînement**:
   - l'algorithme apprend à généraliser
   - construit un modèle prédictif

2. **Test**:
   - à partir du modèle prédictif
   - faire une déduction

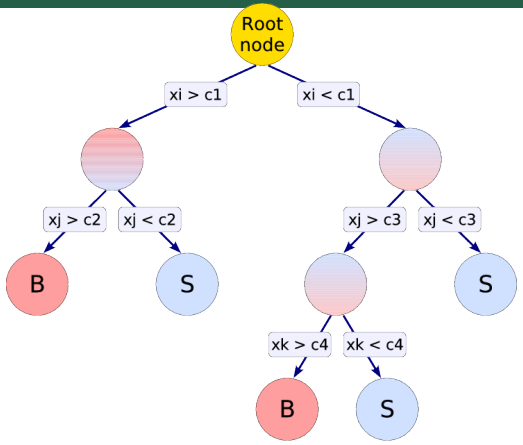
Les algorithmes de classification

De **nombreux algorithmes**

- ⋆ réseaux de neurones
- ⋆ arbres de décision
- ⋆ ...

Et de **nombreux outils**

Arbres de décision



Arbre de décision: succession de coupures séquentielles pour séparer au mieux signal et bruit

- à la **racine** → l'ensemble de l'échantillon
- critère de séparation (**nœuds**) → valeur seuil
- critère d'arrêt
- à chaque **feuille** → type de classe : S ou B

Paramètres ajustables (les plus utilisés)

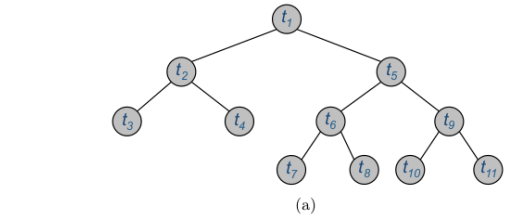
profondeur maximale

nombre minimal d'événements dans les nœuds

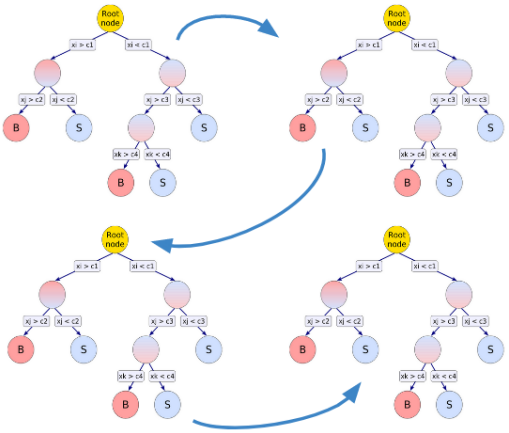
Les arbres de décision: simple, peu de préparation des données,...
sensible au **sur-apprentissage** (**overfitting**)
⟶ éviter le sur-apprentissage: paramètres ajustables, **pruning**

Arbres de décision - **pruning**

- l'arbre de décision maximal (a) est simplifié en un de ses sous-arbres (b), (c)
- en supprimant une ou plusieurs de ses branches de l'arbre non pertinentes
- et en les remplaçant par des feuilles



Arbres de décision - Boosting (BDT)



boosting: construire itérativement de nouveaux arbres

Les événements mal classés (taux d'erreur) sont boostés

Différents algorithmes

- **AdaBoost** (Adaptive Boosting)
- **Gradient Boosting**
  - **XGBoost** (eXtreme Gradient Boosting)

Paramètres ajustables (les plus utilisés)

nombre d'arbres

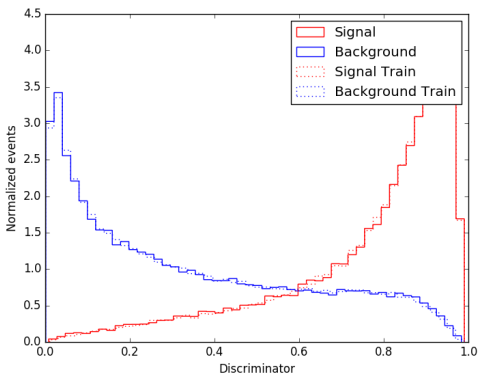
paramètres de boost selon l'algorithme

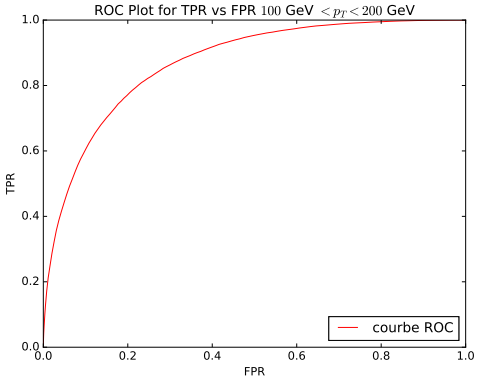
Performances

ROC curve (receiver operating characteristic curve):

graphique montant les performances de l'algorithme de classification

- **TPR** (True Positive Rate) vs **FPR** (False Positive Rate)





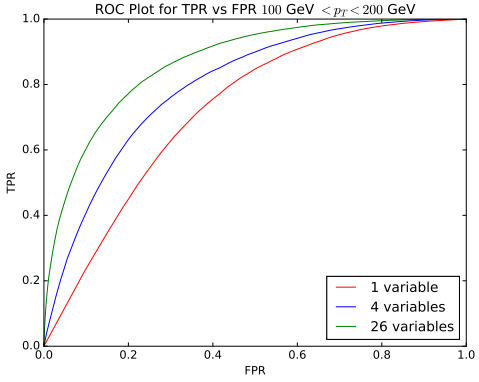
choix du modèle (paramètres), **variables**, **échantillon** (représentatif)

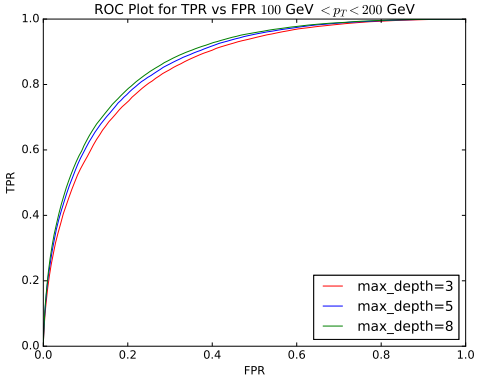
Performances

ROC curve (receiver operating characteristic curve):

graphique montant les performances de l'algorithme de classification

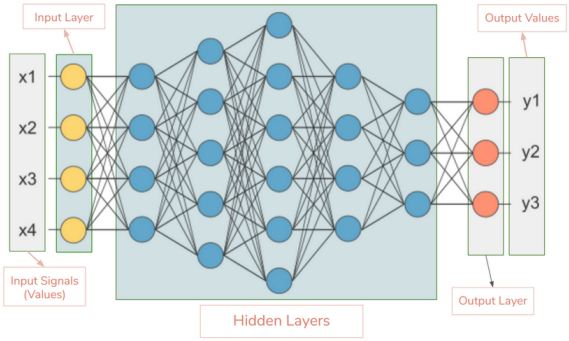
- **TPR** (True Positive Rate) vs **FPR** (False Positive Rate)





choix du modèle (XGBoost), **variables**, **échantillon** (représentatif)

Réseaux de neurones - DNN

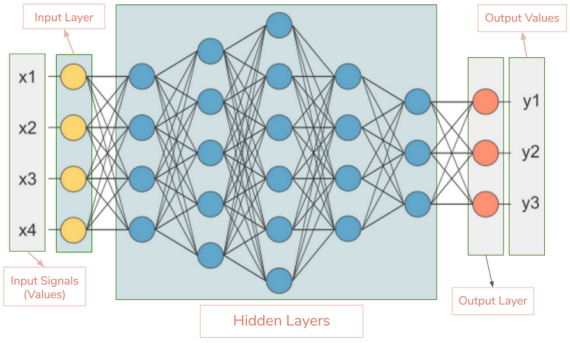


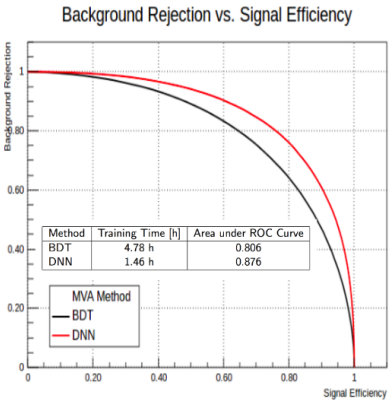
de plus en plus populaire

L'entraînement peut être parallélisé

CPU ou GPU

Réseaux de neurones - DNN





de plus en plus populaire

L'entraînement peut être parallélisé
CPU ou GPU

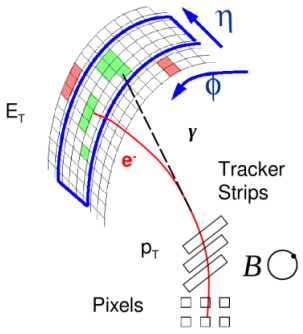
indico.cern.ch/event/798721

Gain en temps d'entraînement

La performance dépend aussi de la paramétrisation du modèle (optimisation)

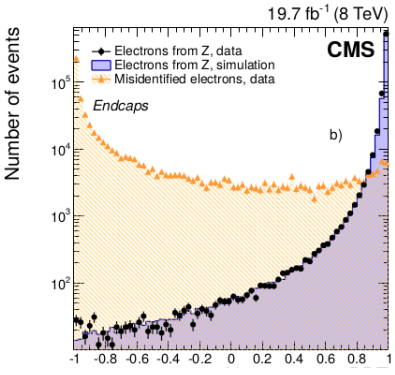
Applications – Identification des électrons

- Un électron est défini à partir de sa **trace**
- ses **dépôts d'énergie** dans différentes cellules du **calorimètre électromagnétique**

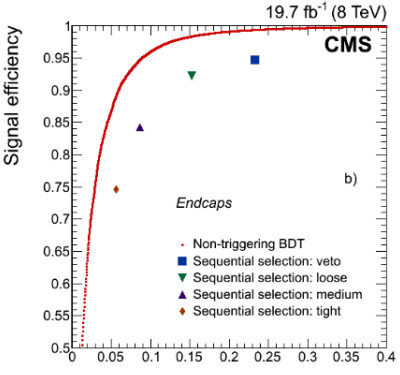


séparer les électrons provenant de sources primaires des autres (bruit de fond)

Applications – Identification des électrons



BDT output



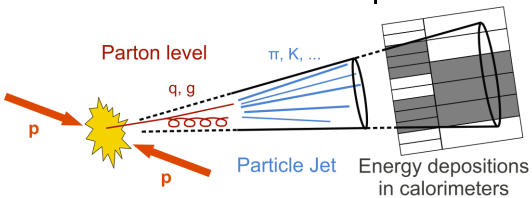
Background efficieny

Amélioration de l'identification avec l'utilisation du BDT.

arxiv:1502.02701

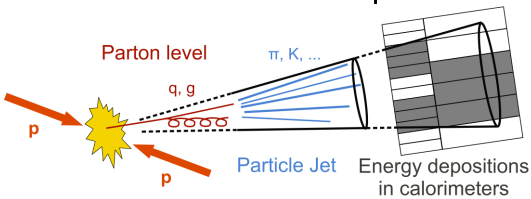
Applications – Identification des jets

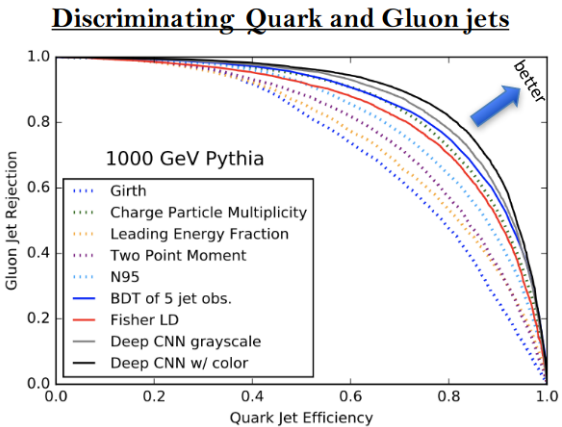
- signature expérimentale laissée par **quarks** (quark-b, quark-c, ⋯) et **gluons**
- ⋆ **traces**
- ⋆ **dépôts d'énergie** calorimètres **EM** et **hadronique**



Applications – Identification des jets

- signature expérimentale laissée par **quarks** (quark-b, quark-c, ···) et **gluons**
- ⋆ **traces**
- ⋆ **dépôts d'énergie** calorimètres **EM** et **hadronique**



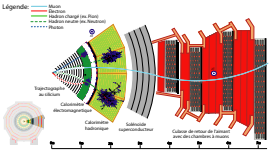


arxiv:1612.01551

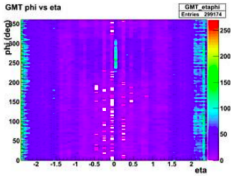
Conclusion

- ML peut intervenir dans plusieurs étapes de reconstruction (Physique des Particules)
- utilisation croissante de ML, surtout amélioration les performances

Identification, reconstruction, analyse



Monitoring



Computing

Électronique