

Методы оптимизации архитектуры и гиперпараметров глубоких нейронных сетей в анализе данных коллайдерных экспериментов

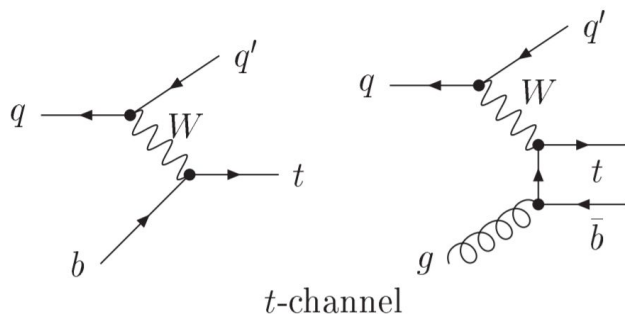
*А. Заборенко, Л. Дудко, П. Волков, Г. Воротников, Э. Абасов, Е. Сивакова, М. Белоброва
НИИЯФ, Физ. ф-т МГУ*

План доклада

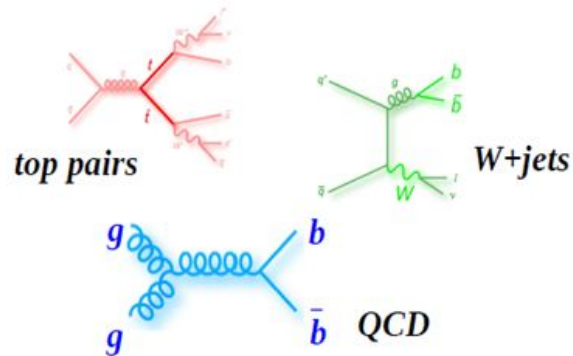
- Физическая задача
- Автоматизация отбора входных переменных
- Пространство гиперпараметров глубокой нейросети:
 - Определение пространства гиперпараметров
 - Оптимизация гиперпараметров
 - Результаты, выводы и рекомендации
- AutoML для физики высоких энергий
- Каскады глубоких нейросетей

Физическая задача

В текущем анализе группы в качестве бенчмарка используется измерение сечения одиночного рождения топ-кварка в эксперименте CMS (LHC).



Фейнмановские диаграммы
одиночного рождения топ-кварка



Фейнмановские диаграммы
фоновых процессов

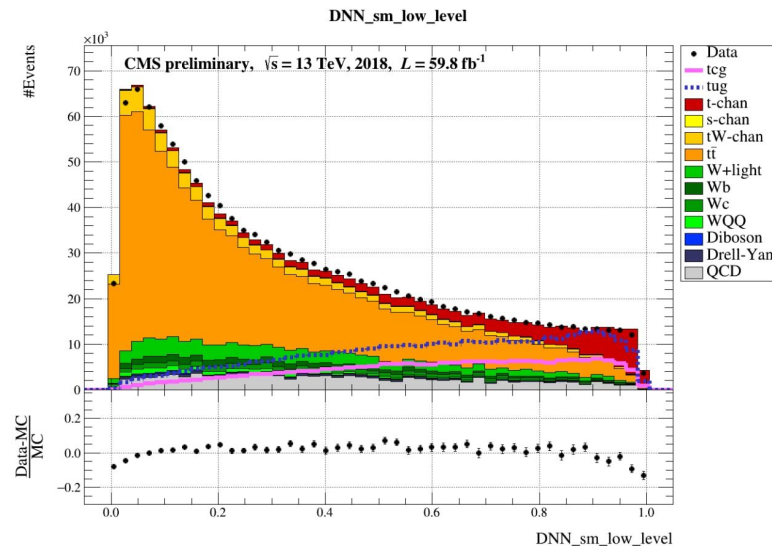
Физическая задача: подавление и разделение процессов

В текущем анализе с помощью глубоких нейросетей (DNN) решаются две основные задачи:

1. Подавление фона многоструйных КХД событий.
2. Разделение процессов. Анализ форм распределений выхода DNN.

Нейронная сеть тренируется так, чтобы выход DNN был в интервале $[0, 1]$, и распределение сигнальных событий группировалось ближе к 1, а событий фона -- к 0.

Используя обрезание по дискриминатору, можно получить набор событий с высокой долей сигнальных событий. Анализируя форму распределений, можно измерить сечения разных процессов.



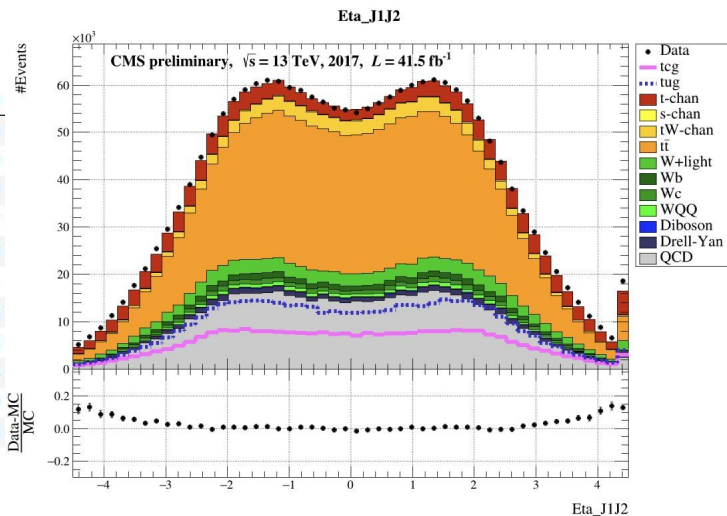
Пример распределения событий для одной из DNN в физическом анализе

Физическая задача с точки зрения машинного обучения

- Датасет из нескольких десятков числовых переменных
- Каждое событие в датасете имеет свой вес, обусловленный физикой исследуемых процессов
- В задаче важна эффективность классификации, а не время вычислений
- В датасете имеется достаточное количество событий для классификации

	log(MtW)	log(MET)	log(Pt_Lep)	DPhi_LepNu	DPhi_WNu	J1_BTag	J2_BTag	LJ_BTag	J1_GTag	J2_GTag	...	log(S_LepJ2)	log(S_LepBJ1)
0	-0.158017	0.471724	1.424701	-1.097130	-0.678345	1.126241	-0.729301	0.273487	-0.634198	1.054127	...	-0.373750	2.240603
1	0.907071	1.969453	2.026695	-0.835542	-0.692116	0.731135	-1.002372	-0.482982	0.311643	-0.385726	...	1.606672	1.845200
2	-0.700078	-3.498501	1.593446	-0.267253	0.588437	-1.064300	1.226518	-0.508284	0.434454	-0.734026	...	0.961636	0.992050
3	-0.123182	-0.294286	-1.153883	-0.220428	-0.304978	0.821729	-0.974507	-0.405791	-0.169487	-0.253373	...	-1.454158	0.462707
4	0.179653	-0.112133	-0.299896	-0.149657	-0.128932	-0.563185	-0.814614	0.037152	0.950013	1.516735	...	0.523618	-0.014386
...	...	...	...	...	...	...	...	...	...	...	...	...	...
182816	-1.748436	-0.260889	-0.977265	-1.402976	-0.953263	-1.055078	1.219687	-0.473089	-0.333425	-0.708669	...	-1.040228	-1.157145
182817	-0.108051	-1.467500	0.093458	0.079289	0.664688	1.145743	-0.989863	-0.445685	-0.696072	-0.149611	...	0.111343	1.032728
182818	0.638017	-0.137699	-0.259352	1.486933	1.281930	-1.030288	1.013197	-0.403290	-0.118881	-0.116235	...	-1.644836	-1.806644
182819	-1.371875	0.690908	-1.158380	-1.384540	-1.003338	-1.030251	1.100544	-0.403185	1.152087	-0.369942	...	-0.116359	-0.164680
182820	-1.937246	-2.537960	-1.094366	-1.161556	-0.597909	1.144940	-0.972643	-0.397793	-0.693653	-0.250767	...	1.358322	1.823278

182821 rows x 72 columns



Выбор входных переменных

Ранее группой был предложен универсальный метод для формирования списка наблюдаемых для нейронных сетей в задачах физики высоких энергий на основе анализа фейнмановских диаграмм.

[International Journal of Modern Physics A Vol. 35, No. 21 (2020) 2050119]

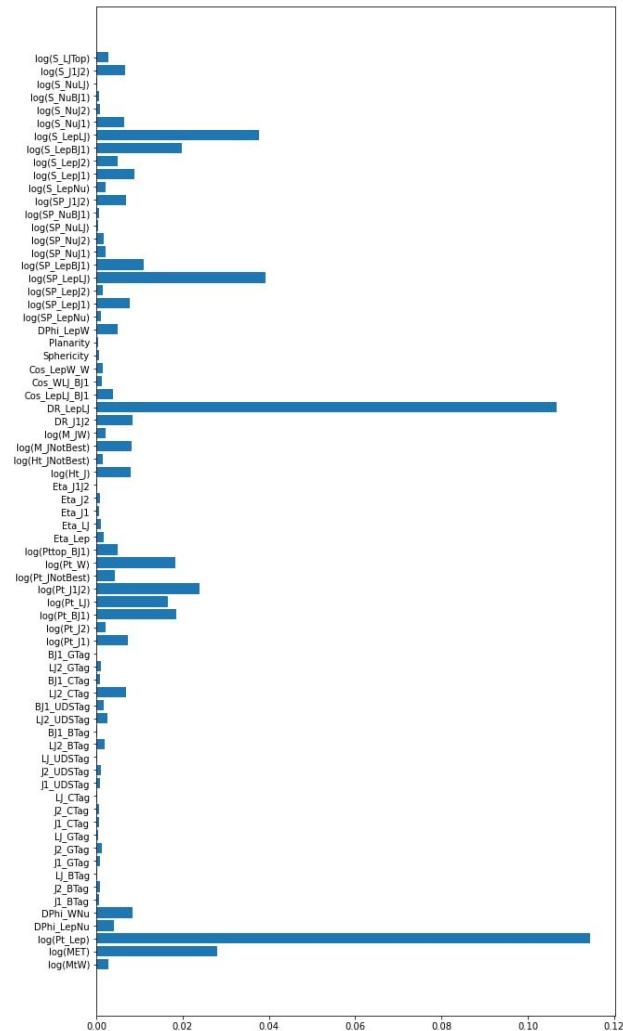
Входные переменные следует нормализовать и применить к ним логарифмическую трансформацию.

Следующим этапом подготовки пространства входных переменных может стать метод `permutation_importance`, реализованный в пакете `sklearn`. Этот метод можно использовать в работе с любыми табличными датасетами и предсказательными моделями.

Выбор входных переменных

В рамках этого метода каждая из переменных “зашумляется”, а затем результат работы модели сравнивается с эталонным. Чем сильнее разница в результатах для отдельно взятой переменной, тем важнее она для текущей классификации.

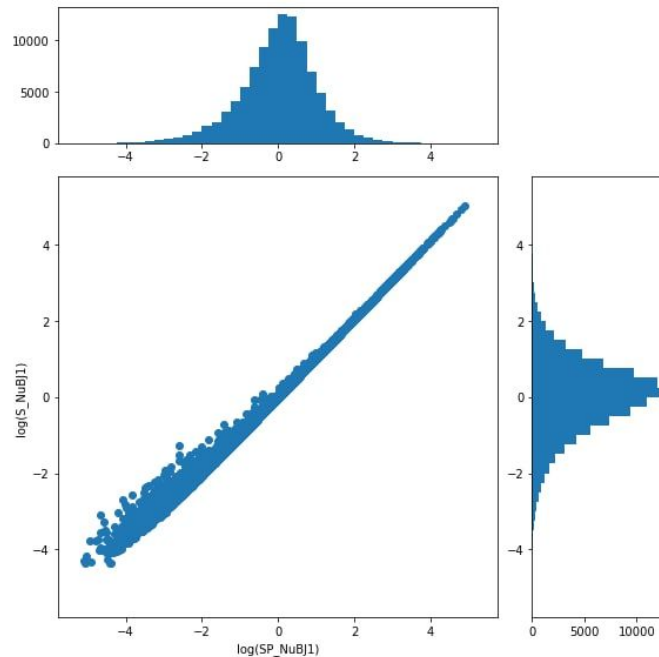
Затем данная операция повторяется несколько раз для всего датасета для уменьшения случайных флуктуаций.



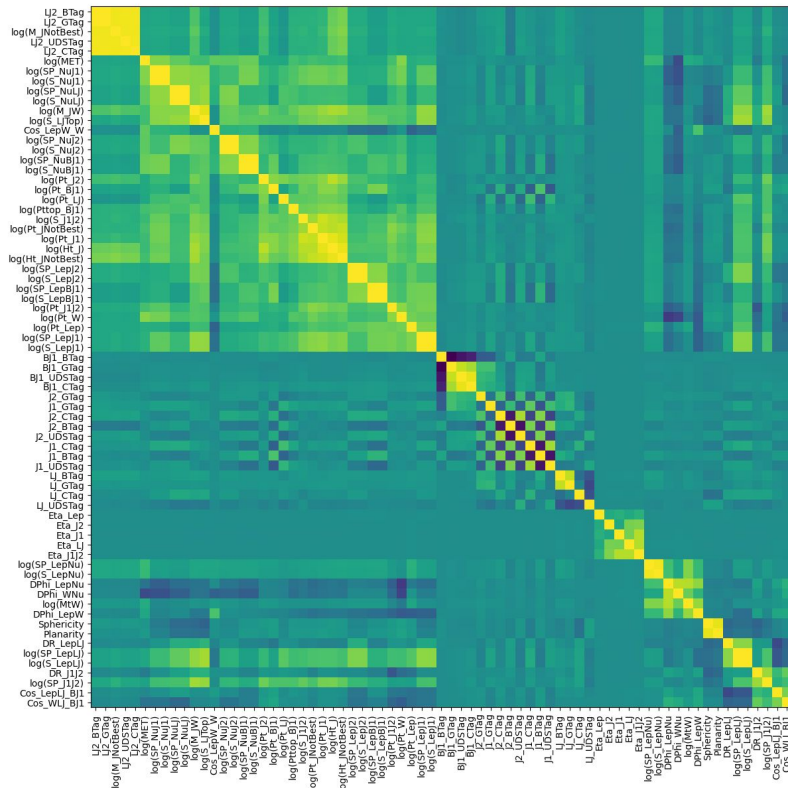
Выбор входных переменных: корреляция

Минус данного метода в том, что если пара переменных скоррелирована и варьируется только одна, то модель может извлечь информацию из незашумленной переменной, уменьшая значимость и той, и другой переменной.

Решением такой проблемы может стать кластеризация переменных на основе их взаимной корреляции.

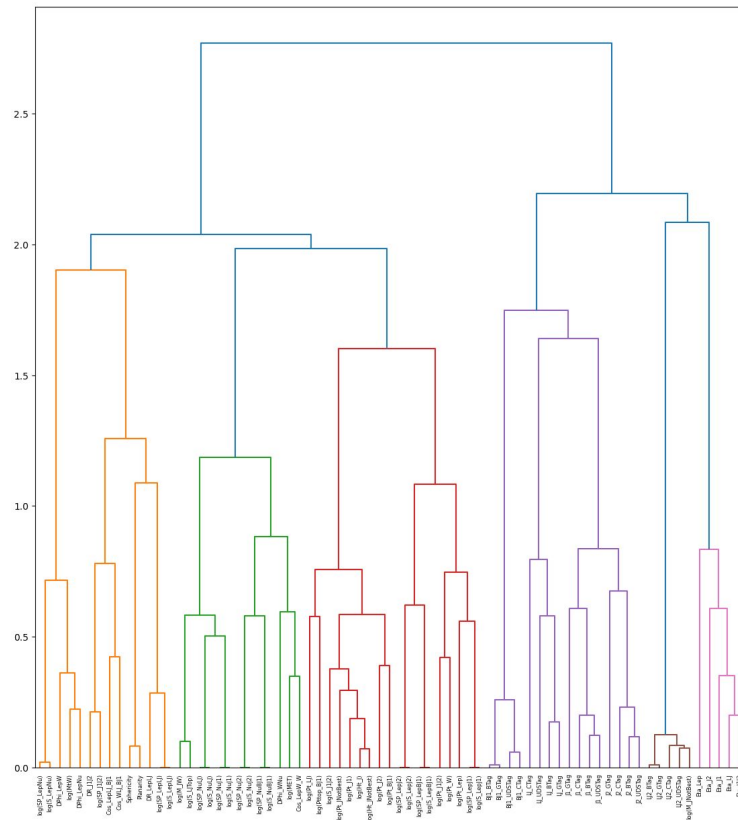


Пример двух сильно скоррелированных переменных. Для классификации выбирается только одна переменная.



Отбор переменных: обрезание по дендрограмме

Различные обрезания по дендрограмме выделяют кластеры переменных, из которых выбирается одна. После этих обрезаний можно провести еще один этап permutation feature importance анализа для получения степени важности уже менее скоррелированных переменных.



Отбор переменных: итоги

Составление оптимального набора переменных для задачи машинного обучения может быть нетривиальной задачей, требующей глубокого понимания датасета. Используемые методы permutation feature importance и кластеризация по корреляции могут помочь исследователю определить, какие переменные являются наиболее информативными для текущей модели.

Именно на эти входные переменные нужно обращать внимание в первую очередь в процессе оптимизации. В текущем анализе описанные методы были применены для формирования конечного набора наблюдаемых.

Пространство гиперпараметров тренировки глубокой сети

Веса глубоких сетей меняются в процессе тренировки, но на эффективность модели влияют и другие параметры, задающие ее конфигурацию. Эти параметры не меняются в процессе градиентного спуска и называются гиперпараметрами.

Физическая задача позволяет использовать полносвязные DNN с прямым распространением сигнала, их основные гиперпараметры следующие:

- Число скрытых слоев
- Ширина скрытых слоев
- Величина дропаута
- Функция активации
- Learning rate
- Константа регуляризации

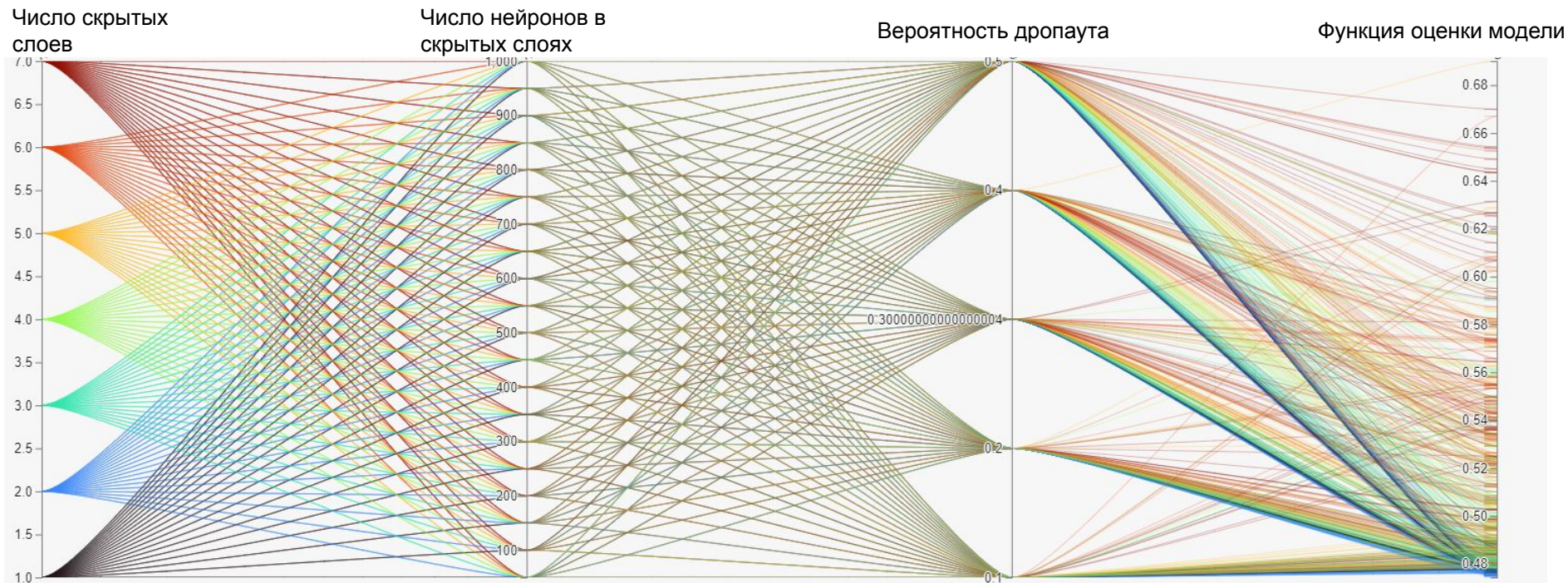
Для контроля переобучения для всех конфигураций используется критерий ранней остановки: сравнение функции ошибки для тренировочного и тестового набора данных.

Для оценки эффективности сравнивается функция ошибки (score), форма распределений выхода сети, ROC AUC.

Оптимизация в пространстве гиперпараметров

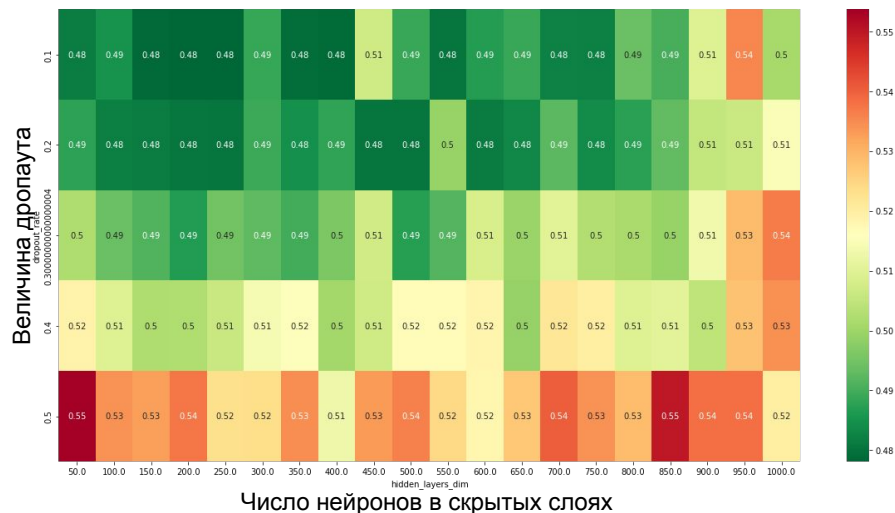
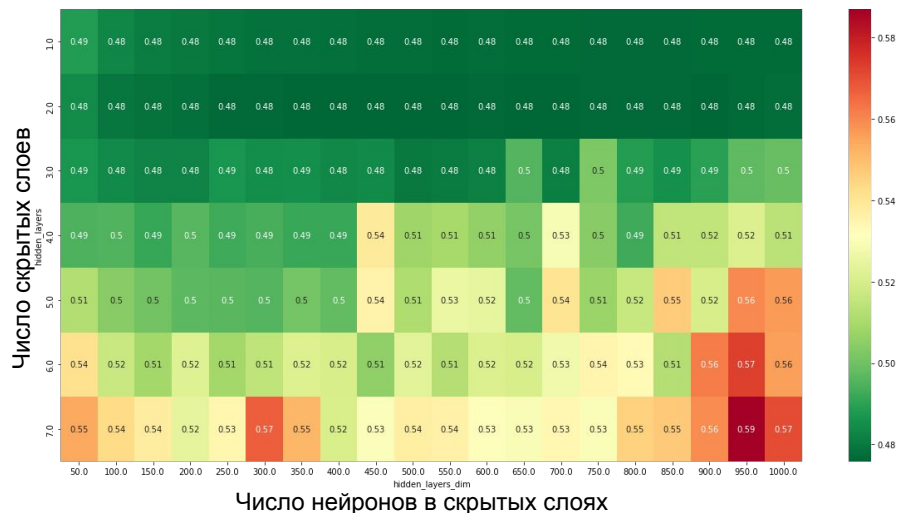
- Grid Search: перебор всех комбинаций подряд, ресурсоемко
- Random Search: случайный перебор комбинаций гиперпараметров, адекватная комбинация находится значительно быстрее, чем перебором
- Hyperband: тренировка различных комбинаций параметров идет небольшое количество эпох, затем результаты после этих эпох сравниваются, и дальнейшее обучение проходят только лучшие комбинации
- Генетические алгоритмы
- Bayesian Optimization: быстрая оптимизация к локальному минимуму
- В принципе, любой оптимизатор на основе машинного обучения

Результаты исследований гиперпространства



Одновременная визуализация трех переменных гиперпространства. Большинство конфигураций показывают стабильную работу: 50% всех комбинаций лежат на расстоянии в 2% от лучшей.

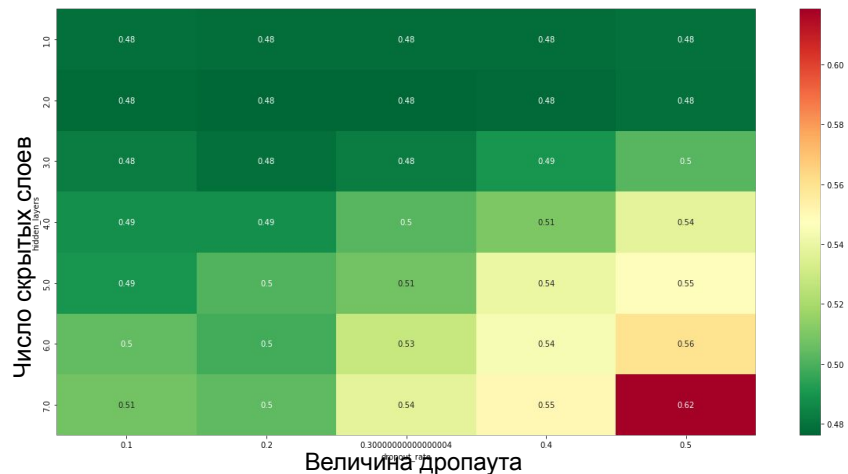
Результаты исследований гиперпространства



Визуализация двумерных зависимостей в пространстве гиперпараметров. Цветом показана средняя разделяющая способность моделей с текущей конфигурацией -- функция бинарной кроссэнтропии (чем меньше, тем лучше).

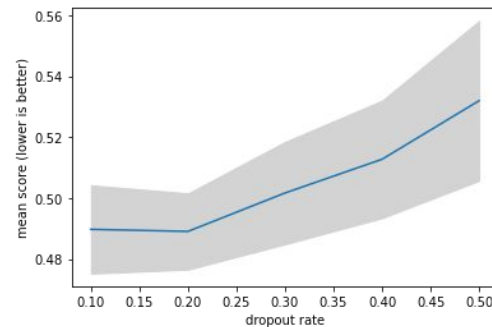
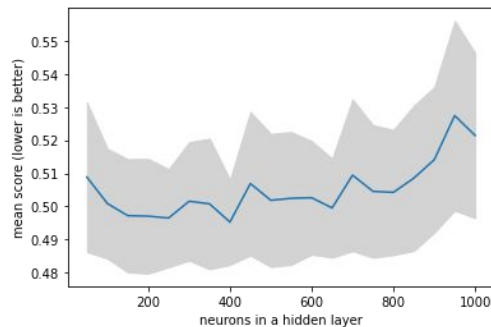
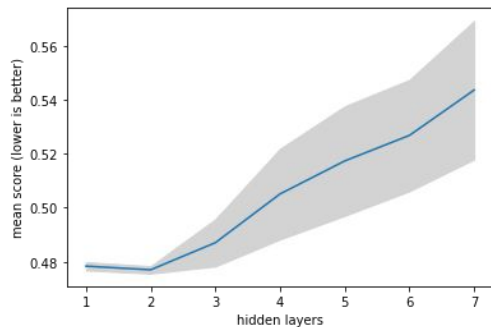
Существует широкая область стабильности для 1-2 слоев и 0.1-0.3 величины дропаута.

Результаты исследований гиперпространства



Слева: зависимость качества модели от числа скрытых слоев и вероятности дропаута. Для 1-2 слоев все величины дропаута приводят к хорошей производительности, а для более глубоких сетей следует уменьшить дропаут до 0.1-0.2.

Внизу: одномерные зависимости качества модели от параметров гиперпространства. По остальным параметрам берется среднее.



Результаты исследований гиперпространства

Итоги:

- Были описаны распространенные методы оптимизации в пространстве гиперпараметров
- Была проведена визуализация пространства гиперпараметров
- Общие выводы для типичной задачи анализа данных коллаидерных экспериментов:
 - модель нейронной сети устойчива ко многим комбинациям гиперпараметров (50% всех комбинаций лежат на расстоянии в 2% от лучшей)
 - для множества вариаций сетей в анализе оптимальный регион гиперпараметров похож: это 1-2 скрытых слоя, 400-800 нейронов в скрытых слоях, 0.2-0.3 вероятность дропаута

AutoML для физики высоких энергий

Методы автоматического машинного обучения направлены на автоматизацию препроцессинга данных, выбора модели, тюнинга ее параметров и ансамблирования нескольких моделей.

Данные физики высоких энергий похожи на табличные данные, поэтому можно использовать программные пакеты, разработанные под этот тип классификационных задач.

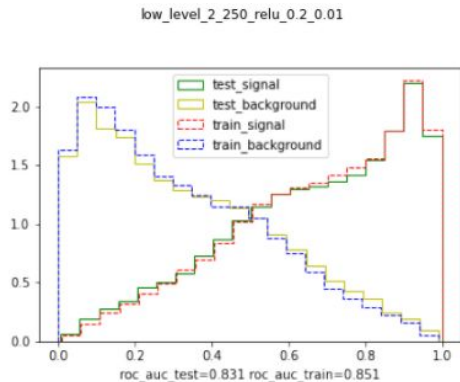
Так как в задачах физики часто важно учитывать веса всех событий, был выбран пакет с поддержкой этого функционала: [mljar-supervised](#).

Использование этого пакета укладывается в две строчки:

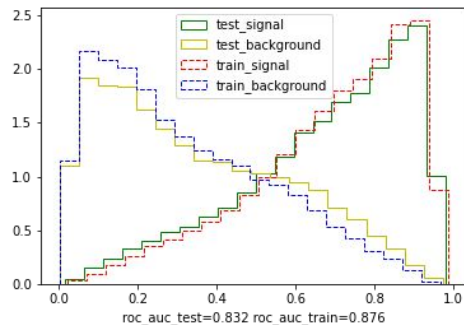
```
automl = AutoML(mode='Optuna', ml_task='binary_classification', results_path=path,  
total_time_limit=2*24*3600, explain_level=2)
```

```
automl.fit(X=features_train, y=labels_train, sample_weight=weights_train)
```

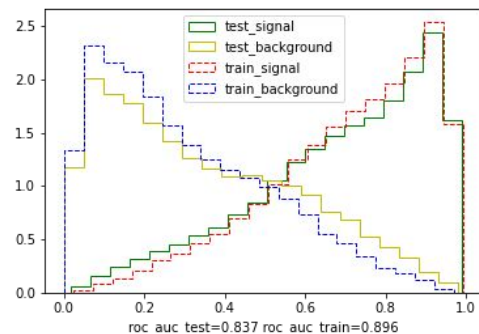
AutoML для физики высоких энергий: сравнение



Нейронная сеть



AutoML (10 минут)



AutoML (2 дня)

В описанном пакете используются следующие алгоритмы и их ансамбли:

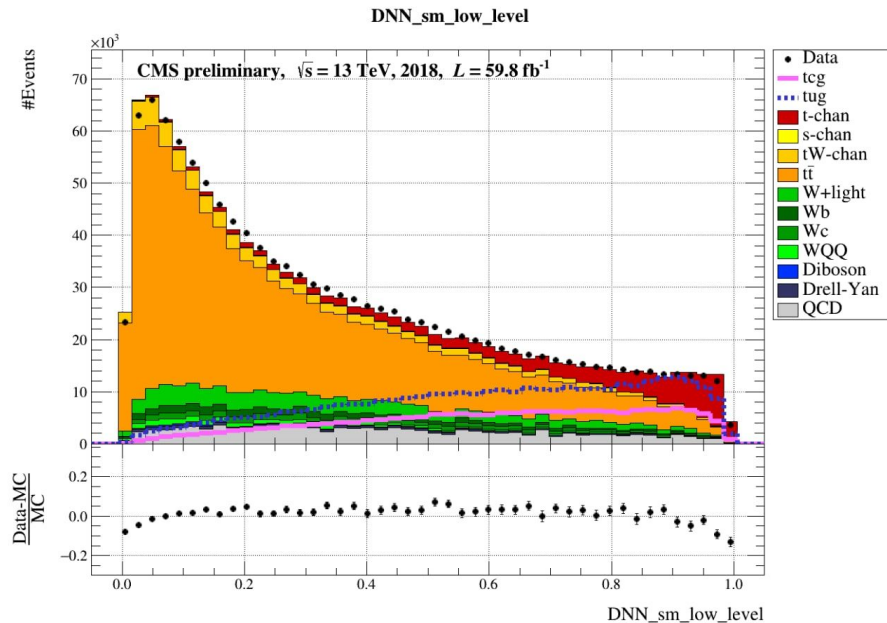
Linear, Random Forest, Extra Trees, LightGBM, Xgboost, CatBoost, sklearn Neural Networks, Nearest Neighbors

Методы AutoML позволили построить компетентную модель за короткое время, но текущее применение AutoML не позволило в достаточной степени контролировать перетренировку.

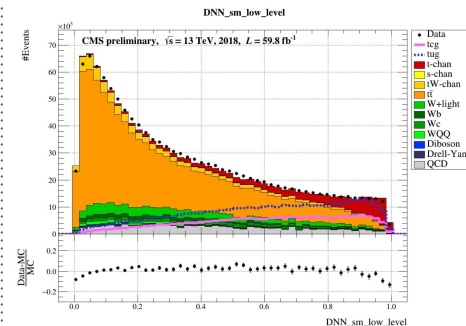
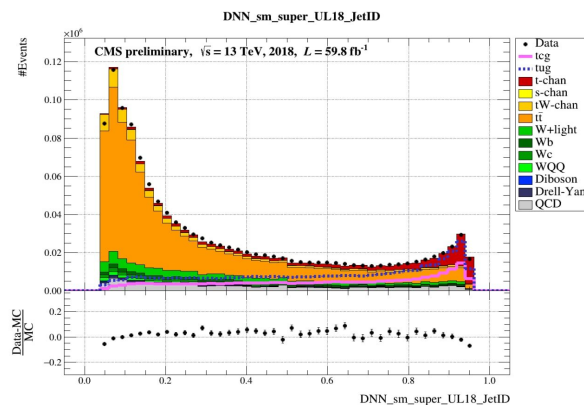
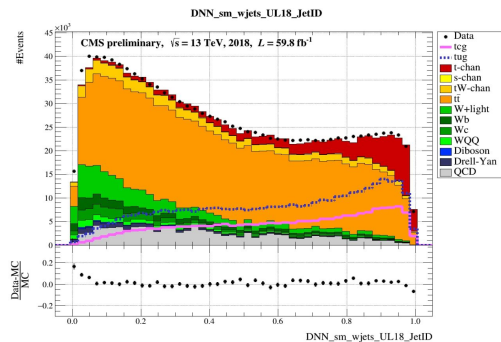
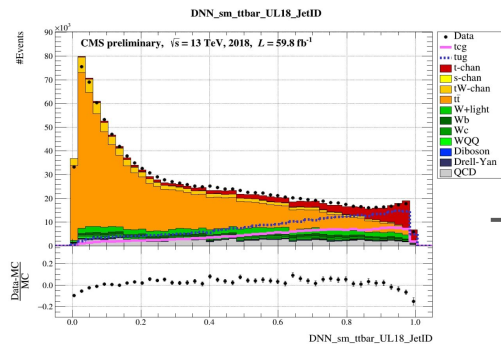
Каскады глубоких нейросетей: идея

В предыдущих итерациях анализа использовалась единая сеть для отделения т-канального процесса от процессов W +jets и t - t bar.

Одна из возможностей дальнейшей оптимизации -- использование каскада нескольких нейросетей: сначала тренируются отдельные сети для разделения т-канала от W +jets и т-канала от t - t bar (так как W +jets и t - t bar сильно отличаются между собой). Затем выходы этих сетей объединяются суперсетью, разделяющей т-канал от W +jets и t - t bar одновременно.



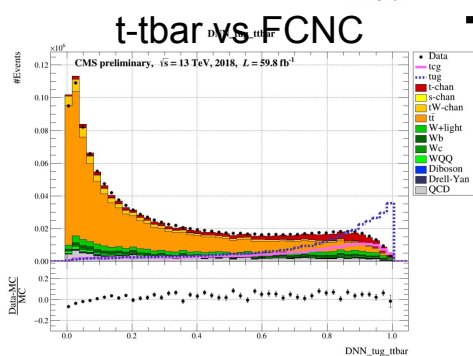
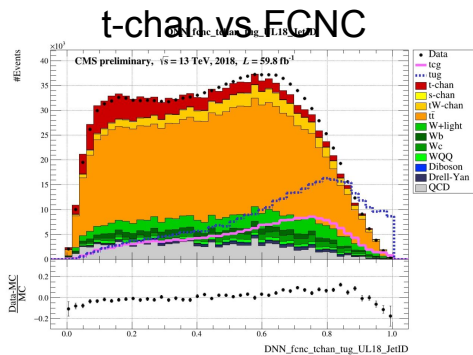
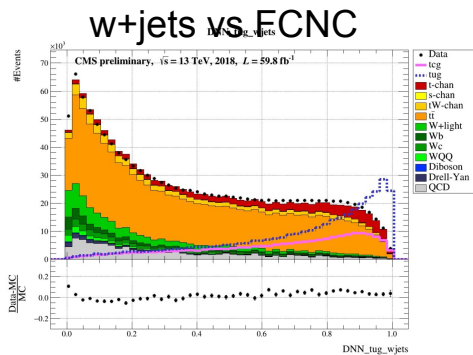
Каскады глубоких нейросетей: применение к СМ



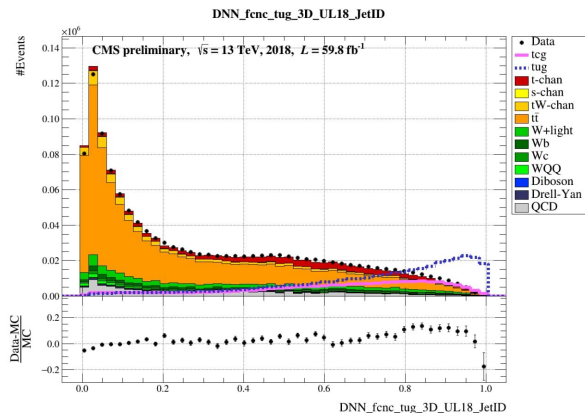
Предыдущая итерация DNN

Дискриминатор полученной суперсети отличается от предыдущей СМ сети лучшим группированием сигнальных событий в сигнальной области и более плоской смешанной областью (0.4 - 0.7)

Каскады глубоких нейросетей: поиск “Новой физики”



+



Для отделения FCNC-tug от SM событий необходимо объединить три сети в одну:

- w+jets vs FCNC
- t-tbar vs FCNC
- t-chan vs FCNC

В результате из-за применения физических закономерностей к обучению нейронных сетей (тренировка против одного физически схожего процесса) было улучшено разделение SM и FCNC событий.

Краткие итоги

- Отбор переменных: один из методов автоматического отбора переменных; анализ важности для модели, анализ корреляций
- Гиперпространство параметров DNN: оптимизация, визуализация, рекомендации
- AutoML в физике: автоматические методы для получения быстрых результатов
- Каскады нейронных сетей: улучшение классификации с помощью применения физических закономерностей к созданию датасетов