Машинное обучение в науках о Земле

Михаил Криницкий

к.т.н., с.н.с.
Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и мониторинга климатических изменений (ЛВОАМКИ)

Интерпретация моделей МО

Михаил Криницкий

к.т.н.
Зав. лабораторией машинного обучения в науках о Земле МФТИ
с.н.с. Института океанологии РАН им. П.П. Ширшова

ПЛАН ЛЕКЦИИ

- Интерпретация линейных моделей
  - Линейная регрессия, логистическая регрессия
  - Неопределенность оценок значимости
- Интерпретация моделей, основанных на деревьях решений
  - Gini impurity или другая функция потерь для деревьев решений
  - Random Forests: оценка значимости на основе значимости отдельных деревьев
- Интерпретация (почти) любых моделей МО
  - деградация качества при перемешивании признака (permutation feature importance)
  - LIME (интерпретация линеаризованной суррогатной модели в окрестности объекта)
- Интерпретация дифференцируемых моделей (напр., ИНС)
  - градиент функции потерь по входным данным
  - связь с методом интерпретации линейных моделей

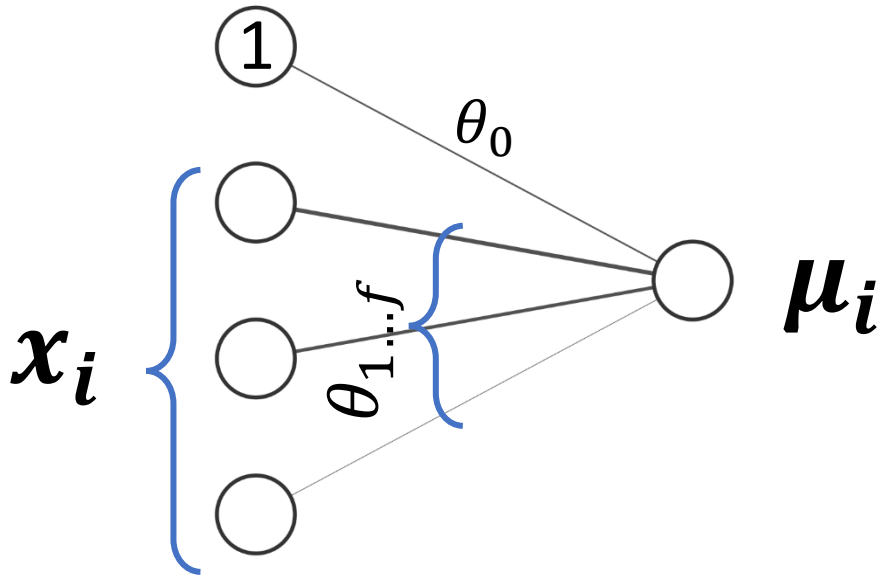
Подход интерпретации моделей МО

- Хочу узнать: оценку значимости признаков
- Оценка вычисляется для обученной модели МО
- Оценка значимости признака – случайная величина, зависит от обучающей выборки (реализации модели МО как «случайной величины»)
    - оценка значимости подчиняется какому-то распределению
    - у оценки значимости есть моменты распределения (среднее, дисперсия), которые можно оценить по выборке (реализаций модели МО, обученных на различных подвыборках), построить доверительные интервалы, etc.
    - выборочные оценки моментов распределений производятся в подходе Bootstrap или скользящего контроля (cross-validation)

Интерпретация линейных моделей

previously on ML4ES

Линейная регрессия



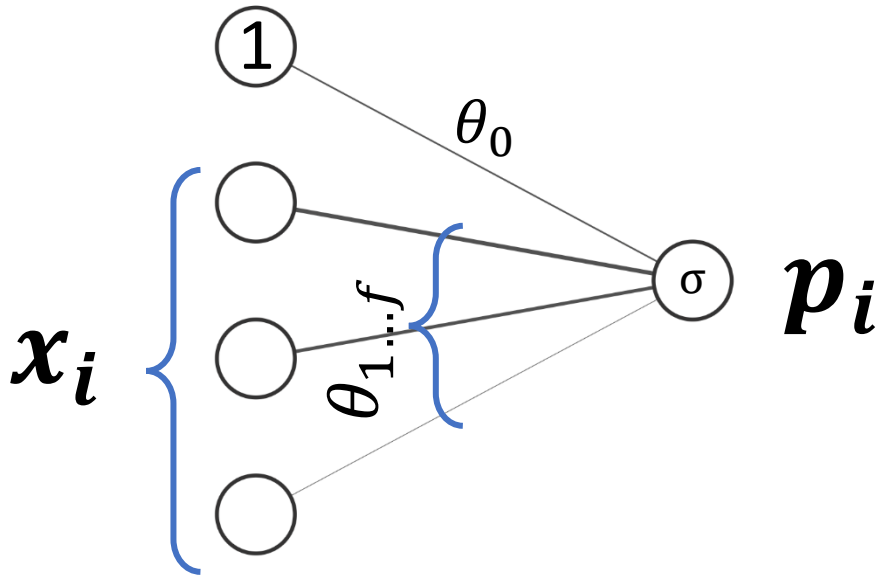
Значимость признаков

- Коэффициенты модели $\boldsymbol{\theta_{1..f}}$
- Сравнение имеет смысл, если признаки одного порядка (напр. стандартизованы)
- Не суммируются в 1
- Не являются «долей объясненной дисперсии»

Интерпретация линейных моделей

previously on ML4ES

Логистическая регрессия



Значимость признаков

- Коэффициенты модели $\boldsymbol{\theta_{1..f}}$
- Сравнение имеет смысл, если признаки одного порядка (напр. стандартизованы)
- Не суммируются в 1
- Не являются «долей объясненной дисперсии»

Интерпретация деревьев решений

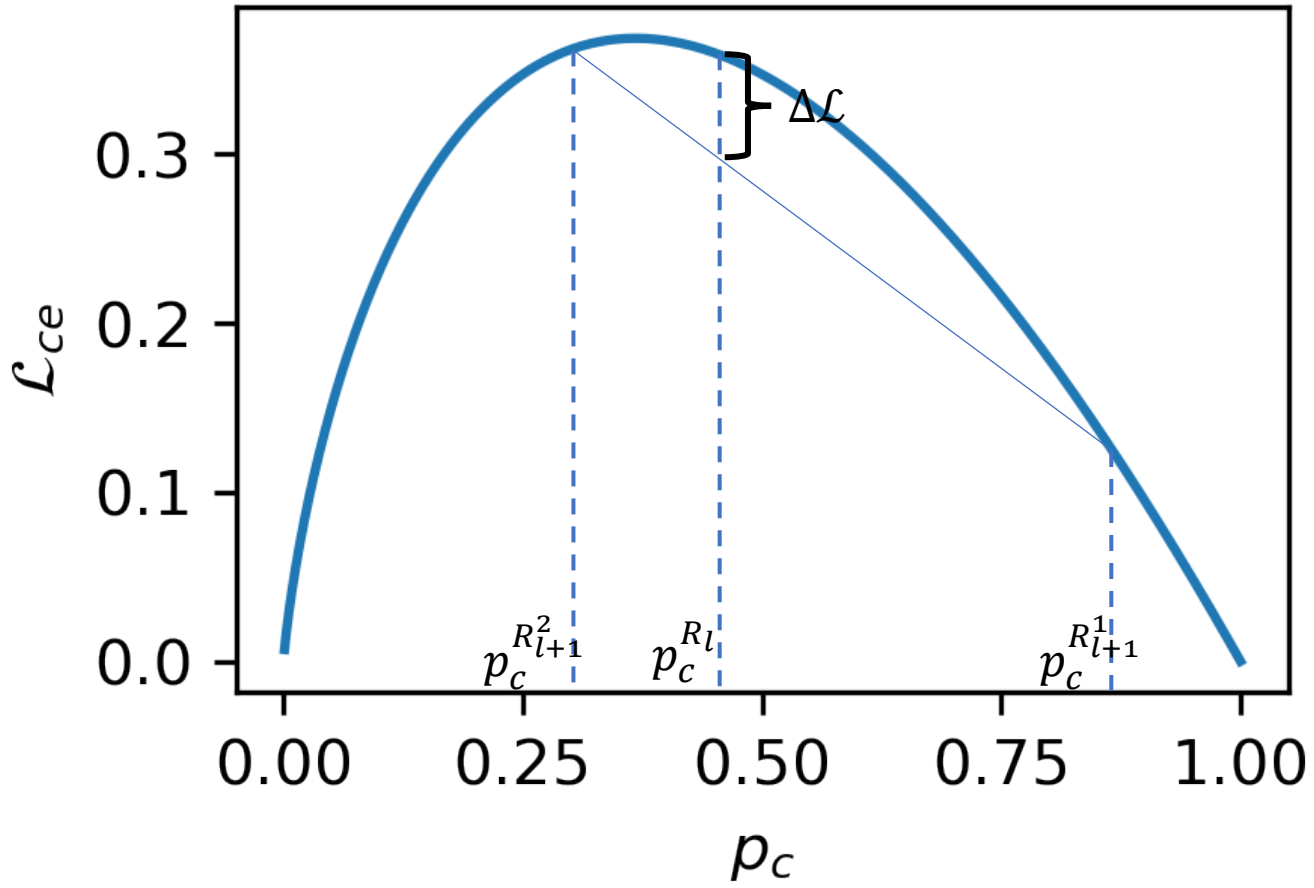
previously on ML4ES

Интерпретация деревьев решений

previously on ML4ES

Разделение обучающей подвыборки R_p^l приводит к тому, что суммарная функция потерь снижается на величину $\Delta\mathcal{L}$.

$$\mathcal{L}_{ce} = -\sum_{c \in \mathbb{Y}} p_c^{(R)} \log p_c^{(R)}$$

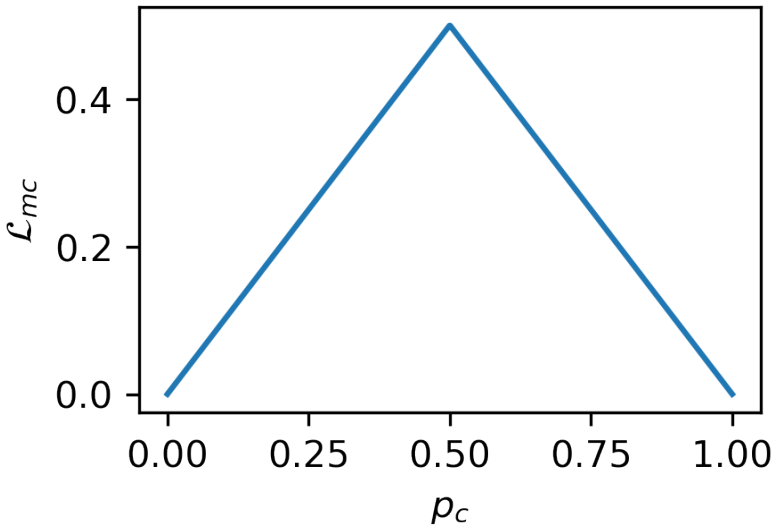


Интерпретация деревьев решений

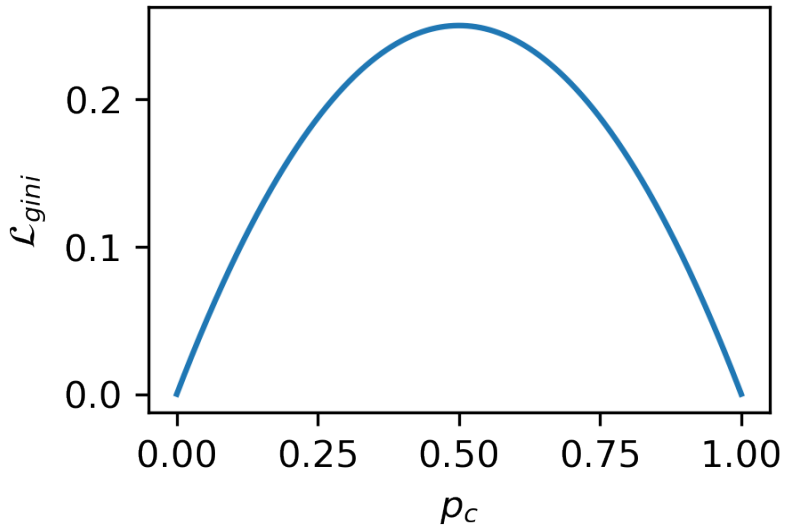
previously on ML4ES

Варианты функции потерь

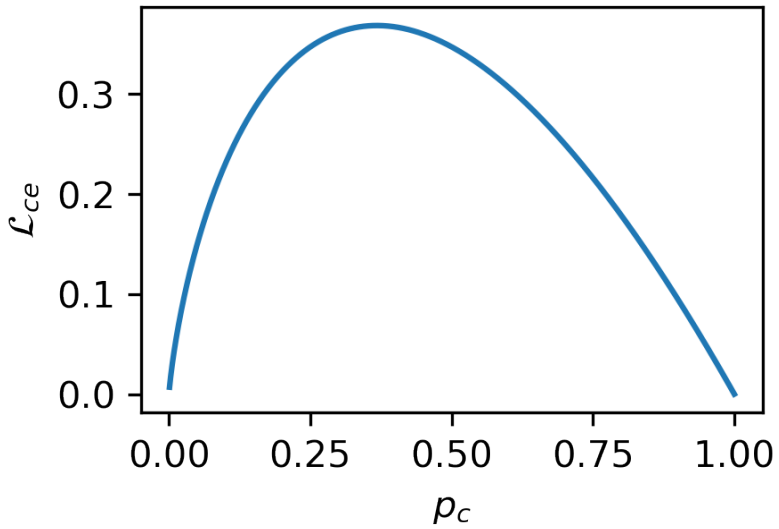
$$\mathcal{L}_{mc} = 1 - \max_{c \in \mathbb{Y}} p_c^{(R)}$$



$$\mathcal{L}_{gini} = \sum_{c \in \mathbb{Y}} p_c^{(R)}(1 - p_c^{(R)})$$



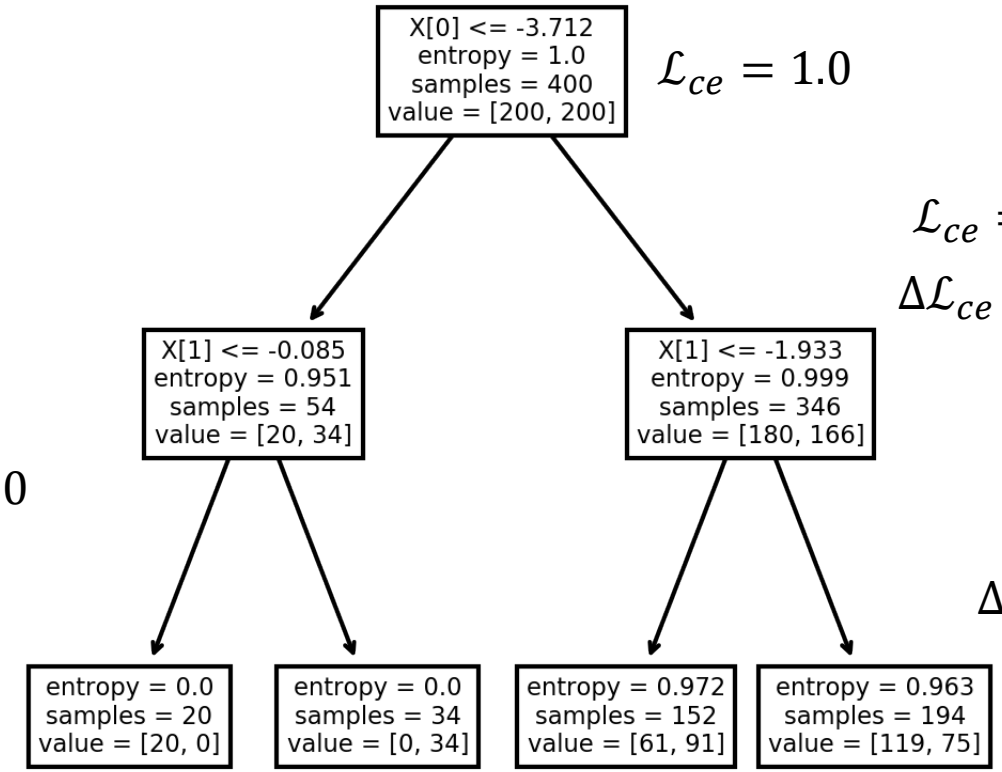
$$\mathcal{L}_{ce} = -\sum_{c \in \mathbb{Y}} p_c^{(R)} \log p_c^{(R)}$$



Интерпретация деревьев решений

previously on ML4ES

Схема ветвления на втором уровне ($l = 1$)



$\mathcal{L}_{ce} = 1.0$

$$\mathcal{L}_{ce} = \frac{54}{400} * 0.951 + \frac{346}{400} * 0.999 = 0.99252$$

$$\Delta\mathcal{L}_{ce} = -0.00748$$

$$\mathcal{L}_{ce} = \frac{20}{54} * 0.0 + \frac{34}{54} * 0.0 = 0.0$$

$$\Delta\mathcal{L}_{ce} = \frac{54}{400} * -0.951 = -0{,}128$$

$$\mathcal{L}_{ce} = \frac{152}{346} * 0.972 + \frac{194}{346} * 0.963 = 0.967$$

$$\Delta\mathcal{L}_{ce} = \frac{346}{400} * -0.0256 = -0{,}0221148$$

Интерпретация деревьев решений

Значимость признаков

- (нормализованный) накопленный вклад признака в снижение функции потерь
- Нет необходимости нормализовать сами признаки
- Значимости суммируются в 1
- Отражает вклад в снижение функции потерь (не в повышение меры качества)
- Обладает свойством смещенности значений значимости к признакам с бОльшей мощностью множества значений (напр., действительным признакам)

sklearn.tree.DecisionTreeClassifier
sklearn.tree.DecisionTreeRegressor
} model.feature_importances_

Интерпретация Random Forests

previously on ML4ES

Bagging эксплуатирует подход обучения большого количества ($K \gg 1$) моделей, склонных к переобучению (σ^2 - существенна, но ρ сильно меньше единицы, алгоритмы раскоррелированы за счет склонности к переобучению и за счет обучения на различающихся подвыборках).

Способ применения в случае решающих деревьев: обучить очень много довольно решающих деревьев до конца (не ограничивая их глубину, без регуляризаций); обучать на bootstrap-выборках, **агрегировать результаты по принципу простого голосования** (в случае классификции) или **простого осреднения** (в случае регрессии).

Интерпретация Random Forests

Значимость признаков MDI (Mean Decrease in Impurity)

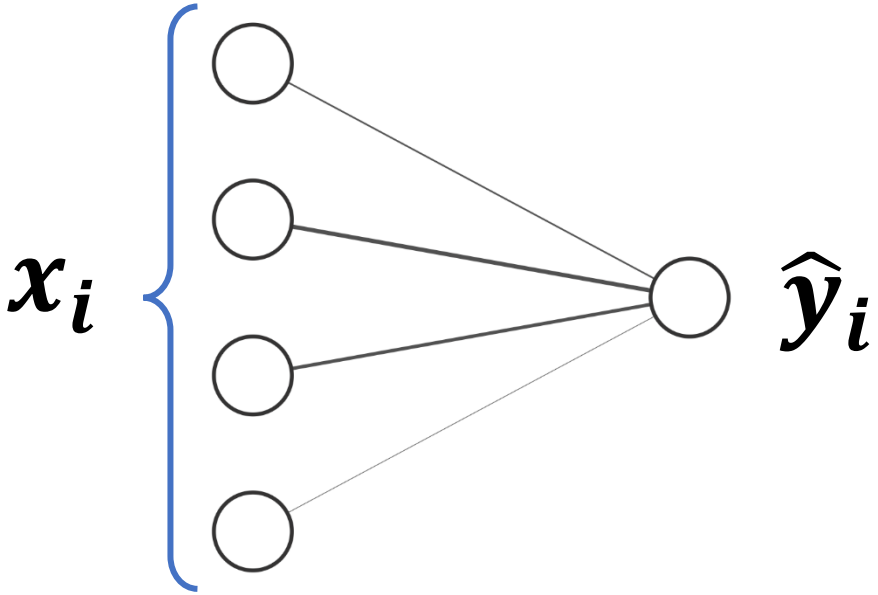
- осредненное значение значимостей признака, оцененных по членам композиции (отдельным деревьям)
- Нет необходимости нормализовывать признаки
- Суммируются в 1
- Отражают вклад в снижение функции потерь (не в повышение меры качества)
- Обладает свойством смещенности значений значимости к признакам с бОльшей мощностью множества значений (напр., действительным признакам)

sklearn.ensemble.RandomForestRegressor
sklearn.ensemble.RandomForestClassifier
} model.feature_importances_

Интерпретация произвольных моделей МО

sklearn.inspection.permutation_importance

- Значимость признака - мера деградации качества при случайном перемешивании значений признака*



Интерпретация произвольных моделей МО

sklearn.inspection.permutation_importance

- Значимость признака - мера деградации качества при случайном перемешивании значений признака*
- мера качества - любая, задаваемая исследователем;
- оценивается для **обученной** модели в **режиме применения**, **на ВСЕЙ тестовой выборке;**
- нет необходимости нормализовывать сами признаки;
- оценки значимости не нормированы (нужно нормировать вручную);
- отражают дифференциальную деградацию меры качества;
- неопределенность значимости оценивают в подходе кросс-валидации или bootstrap-сэмплирования

Интерпретация произвольных моделей МО

LIME* (Local Interpretable Model-agnostic Explanations)

- Значимость признака оценивается как значимость этого же признака для суррогатной модели, обученной на результатах сложной модели в окрестности объекта;
- Может применяться для любой обученной модели МО в окрестности тестовых объектов;
- Модель для регрессии – линейная регрессия, модель для классификации – логистическая регрессия (бинарная/мультиномиальная); иногда – деревья решений;
- Для порождения обучающей выборки для суррогатной модели в признаки вносится шум**, позволяющий сэмплировать синтетические объекты в окрестности (одного) тестового;
- При обучении суррогатной модели синтетическим объектам назначается вес в пространстве признаков, зависящий от расстояния до оригинального тестового примера

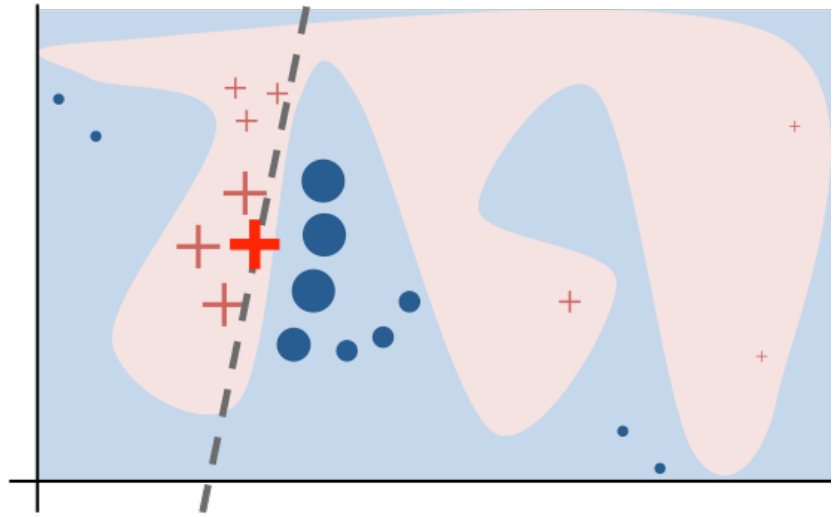
*Ribeiro M. T., Singh S., Guestrin C. «Why Should I Trust You?»: Explaining the Predictions of Any Classifier // 2016

** характер шума зависит от типа данных

Интерпретация произвольных моделей МО

LIME* (Local Interpretable Model-agnostic Explanations)

- LIME: много гиперпараметров:
  - способ сэмплирования объектов для обучения суррогатной модели;
  - способ вычисления весов синтетических объектов;
  - гиперпараметры суррогатной модели (вид, способ оценки значимости признаков).



```
$ pip install lime
```

https://github.com/marcotcr/lime

*Ribeiro M. T., Singh S., Guestrin C. «Why Should I Trust You?»: Explaining the Predictions of Any Classifier // 2016

** характер шума зависит от типа данных

Интерпретация нейросетей

- Значимость признака оценивается как относительная оценка магнитуды градиента функции потерь или целевой переменной по признакам в окрестности тестового объекта;
- может применяться для любой обученной нейросети в окрестности тестовых объектов;
- в случае «линейной однослойной нейросети» такая оценка эквивалентна оценке по коэффициентам линейной модели.