Отношение сигнал/шум в астрофизических наблюдениях

Лекция 15

Современные методы оценки S/N в астрофизических наблюдениях и их применение в исследованиях. Автоматизированный анализ, статистическое моделирование и методы машинного обучения

КазНУ имени аль-Фараби

Цель и план лекции

Цель

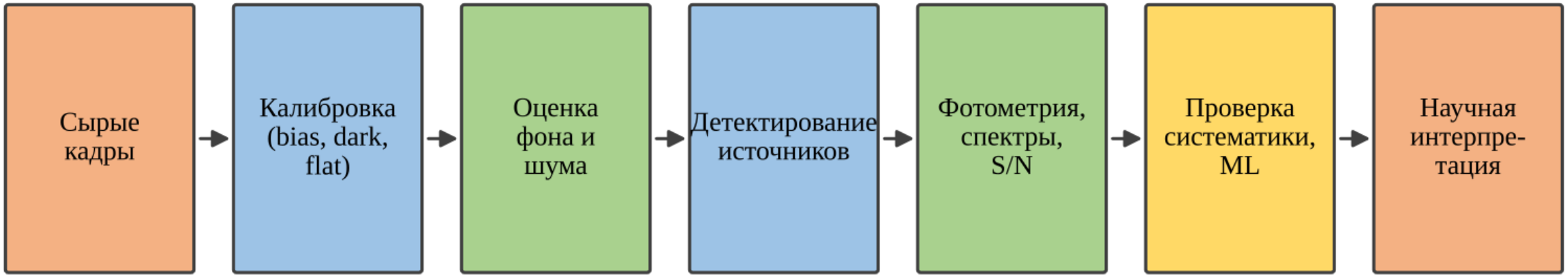
Обобщение изученных методов, рассмотрение автоматизированных процедур анализа, статистического моделирования и машинного обучения, а также формирование подхода к корректной оценке S/N в научной работе.

1. Введение
2. Автоматизированные конвейеры обработки
3. Статистическое моделирование и байесовские методы
4. Методы машинного обучения
5. Рекомендации по корректной оценке S/N
6. Заключение курса
7. Современные приборы и задачи анализа

1 Введение

- Современные обзоры производят огромные объёмы данных: обсерватория Веры Рубин будет ежедневно регистрировать терабайты изображений и формировать миллионы оповещений, а Gaia, TESS, JWST и радиотелескопы создают большие каталоги.

1 Введение (продолжение)



Последовательность автоматизированной обработки астрономических данных

2 Автоматизированные конвейеры обработки

Конвейер обработки данных включает несколько этапов.

Первичная калибровка

Вычитание смещения и темнового тока, деление на плоское поле, маскирование дефектных пикселей и следов космических лучей.

Оценка фона и шума

Построение карт фона и карт дисперсии или веса для каждого изображения. Отсутствие корректных карт приводит к ошибочным значениям S/N во всех последующих каталогах.

Астрометрия и фотометрическая привязка

Определение положения и нуль-пунктов.

Пакеты, такие как Astropy, photutils, ccdproc, образуют открытую экосистему для обработки данных на языке Python и обеспечивают воспроизводимость вычислений.

2 Автоматизированные конвейеры обработки (продолжение)

Обнаружение и измерение

Создание каталога программами SExtractor, photutils и другими пакетами.

Контроль качества

Автоматические проверки, например статистика FWHM, эффективная глубина кадра, сравнение со стандартными каталогами.

Пакеты, такие как Astropy, photutils, ccdproc, образуют открытую экосистему для обработки данных на языке Python и обеспечивают воспроизводимость вычислений.

3 Статистическое моделирование и байесовские методы

- Байесовская иерархическая модель позволяет определить совместное распределение параметров объектов и популяции, учитывая погрешности измерений и порог отбора.
- При прямой подгонке модели к данным (forward modeling) рассчитывается предсказание для пикселей с учётом PSF, детектора и шума.

4 Методы машинного обучения

Классификация реальных и ложных транзиентов

Модели обучаются на искусственных и подтверждённых объектах и отбрасывают артефакты разностных изображений, что значительно уменьшает нагрузку на исследователей.

Поиск транзитов

Шэллу и Вандербург (2018) применили свёрточную сеть для классификации сигналов Kepler.

Подавление шума

Сети типа U-Net, метод Noise2Noise (Лехтинен и др., 2018) обучаются на парах независимых шумных изображений; метод GAN применялся для восстановления изображений галактик.

Основные риски состоят в следующем. Модель может «галлюцинировать» структуры, отсутствующие в данных.

4 Методы машинного обучения (продолжение)

Оценка фона и характеристик шума

Модели предсказывают уровень шума и наличие артефактов на основе признаков.

Выбор наблюдений

Активное обучение и последовательный анализ помогают определить наиболее информативные наблюдения.

Основные риски состоят в следующем. Модель может «галлюцинировать» структуры, отсутствующие в данных.

5 Рекомендации по корректной оценке S/N

Определение

Необходимо чётко указывать, что понимается под S/N, для какой области и какого элемента.

Шумовая модель

Следует указывать источники шума и способ их определения, а также проверять независимость и нормальность остатков.

Систематика

Следует отдельно оценивать систематические ошибки.

Проверки

Следует использовать инъекции, повторные выборки и независимые инструменты.

Воспроизводимость

Следует публиковать код и исходные данные, если это возможно.

Значимость

Необходимо учитывать количество выполненных проверок.

5 Автоматическое обнаружение источников

Листинг 15.1.

```
import numpy as np
from scipy import ndimage as ndi

rng = np.random.default_rng(15)
n, sigp, sky, rn = 400, 2.0, 100.0, 5.0
true = rng.integers(20, n - 20, (40, 2))
flux = rng.uniform(150, 3000, 40)

y, x = np.indices((n, n))
img = np.full((n, n), sky, float)
for (cy, cx), F in zip(true, flux):
    img += F / (2 * np.pi * sigp**2) * np.exp(
        -((x - cx)**2 + (y - cy)**2) / (2 * sigp**2))
data = rng.poisson(img) + rng.normal(0, rn, (n, n))

bkg = np.median(data)
noise = 1.4826 * np.median(np.abs(data - bkg))
smooth = ndi.gaussian_filter(data - bkg, sigp)
sn_map = smooth / (noise * np.sqrt(1 / (4 * np.pi * sigp**2)))
label, nobj = ndi.label(sn_map > 5)
pos = np.array(ndi.maximum_position(sn_map, label,
                                    range(1, nobj + 1)))
```

5 Автоматическое обнаружение источников (продолжение)

Листинг 15.1.

```
d = np.hypot(*(pos[:, None] - true[None, :]).transpose(2, 0, 1))
print(f"Фон: {bkg:.1f}, шум на пиксель: {noise:.2f}")
print(f"Обнаружено: {nobj}, из них совпало: "
      f"{(d.min(1) < 3).sum()}")
print(f"Найдено заданных: {(d.min(0) < 3).sum()} из {len(true)}")
```

5 Результат: Автоматическое обнаружение источников

```
Фон: 100.1, шум на пиксель: 11.27
Обнаружено: 37, из них совпало: 36
Найдено заданных: 37 из 40
```

5 Отделение источников от артефактов с помощью ML

Листинг 15.2.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_auc_score
from sklearn.model_selection import train_test_split

def sample(kind):
    a = rng.uniform(4, 9)             # кажущийся S/N
    w = rng.uniform(1.6, 2.2)
    if kind == 0:
        w = rng.uniform(0.4, 1.0)     # узкие артефакты
    h = rng.uniform(.9, 1.1) if kind else rng.uniform(.2, 2.0)
    return [a, w, h]

X = np.array([sample(k) for k in range(2) for _ in range(1500)])
Y = np.repeat([0, 1], 1500)
Xa, Xb, ya, yb = train_test_split(X, Y, test_size=0.4,
                                  random_state=1, stratify=Y)
clf = RandomForestClassifier(200, random_state=1).fit(Xa, ya)
print(f"AUC случайного леса: "
      f"{roc_auc_score(yb, clf.predict_proba(Xb)[:, 1]):.3f}")
print(f"AUC порога по S/N:   {roc_auc_score(yb, Xb[:, 0]):.3f}")
print("Важность признаков (S/N, ширина, эллиптичность):",
      np.round(clf.feature_importances_, 3))
```

5 Результат: Отделение источников от артефактов с помощью ML

```
AUC случайного леса: 1.000
AUC порога по S/N:   0.484
Важность признаков (S/N, ширина, эллиптичность): [0.008 0.735 0.258]
```

В этом модельном примере порог только по S/N не различает реальные объекты и артефакты, а использование формы объекта разделяет эти два класса.

6 Заключение курса

- В рамках курса последовательно рассмотрены физические и статистические основы S/N, основные шумы детектора и атмосферы, фотометрические и спектроскопические наблюдения, обнаружение слабых сигналов, оптимальное планирование и объединение данных, анализ временных рядов и коррелированных шумов.

7 Современные приборы и задачи анализа

- Для разных областей астрофизики характерны свои технологические тенденции. Обзор LSST будет сочетать большую глубину с многократным наблюдением одного и того же участка неба, поэтому особую роль приобретут вероятностные каталоги, в которых для объектов указываются функции вероятности, а не единственные значения параметров.

- В анализе изображений всё более важную роль приобретает восстановление функции рассеяния и моделирование с использованием дифференцируемых программ.
- Вероятностное программирование делает возможным построение сложных иерархических моделей с помощью таких инструментов, как PyMC, Stan и NumPyro.
- Необходимо уделять внимание калибровке неопределённостей моделей машинного обучения.

7 Современные приборы и задачи анализа (продолжение)

- В заключение отметим, что независимо от выбранного инструмента основой остаётся физически обоснованная модель шума.
- Упражнение: подготовьте для итогового семинара краткий анализ S/N реальных данных своей диссертационной работы, включающий бюджет шума, сравнение расчётной и измеренной точности и обсуждение основных ограничений.

8 Разбор примера

1. На матрице 4096×4096 пикселей при области поиска 10 пикселей имеется примерно $1{,}68 \cdot 10^6$ независимых проверок.
2. Если модель ML пропускает ложные объекты с вероятностью 10^{-3}, а исходный детектор при 3σ создаёт 2265 ложных объектов, то после отбора останется около 2.
3. При этом важна доля сохранённых реальных объектов: если она составляет 90 %, то 10 % реальных объектов будут потеряны, и это необходимо учитывать при определении полноты каталога.

9 Вопросы для обсуждения на занятии

Применение

Какая часть процедуры в ваших диссертационных исследованиях может быть автоматизирована?

Риски

Как проверить, что шумоподавление с использованием нейросети не создало ложных структур?

Оценка

Какие сведения должны быть приведены в статье для обоснования значения S/N?

Контрольные вопросы

1. Какие этапы включает автоматический конвейер обработки?
2. Для чего требуются карты дисперсии изображений?
3. В чём преимущество иерархических байесовских моделей?
4. Какие задачи решаются с помощью методов машинного обучения?
5. Почему S/N после нейросетевого шумоподавления необходимо интерпретировать осторожно?
6. Как применять искусственные источники для проверки результатов?
7. Какие рекомендации важно соблюдать при оценке S/N в публикациях?
8. Как значимость обнаружения связана с числом проверок?

Список литературы

1. Ivezić Ž., Connolly A.J., VanderPlas J.T., Gray A. Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. – Updated ed. – Princeton: Princeton University Press, 2020.

2. Astropy Collaboration The Astropy Project: Sustaining and growing a community-oriented open-source project and the latest major release (v5.0) of the core package // The Astrophysical Journal. – 2022. – Vol. 935. – Art. 167.

3. Bertin E., Arnouts S. SExtractor: Software for source extraction // Astronomy and Astrophysics Supplement Series. – 1996. – Vol. 117. – P. 393–404.

4. Alard C., Lupton R.H. A method for optimal image subtraction // The Astrophysical Journal. – 1998. – Vol. 503. – P. 325–331.

5. Zackay B., Ofek E.O., Gal-Yam A. Proper image subtraction: Optimal transient detection, photometry, and hypothesis testing // The Astrophysical Journal. – 2016. – Vol. 830. – Art. 27.

6. Shallue C.J., Vanderburg A. Identifying exoplanets with deep learning: A five-planet resonant chain around Kepler-80 and an eighth planet around Kepler-90 // The Astronomical Journal. – 2018. – Vol. 155. – Art. 94.

7. Lehtinen J., et al. Noise2Noise: Learning image restoration without clean data // Proc. 35th International Conference on Machine Learning: PMLR. – 2018. – Vol. 80. – P. 2965–2974.

Список литературы (продолжение)

8. Schawinski K., Zhang C., Zhang H., Fowler L., Santhanam G.K. Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit // Monthly Notices of the Royal Astronomical Society. – 2017. – Vol. 467. – P. L110–L114.

9. Pedregosa F., et al. Scikit-learn: Machine learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.