Что такое Statistical Learning

Антон Коробейников

5 сентября 2014 г.

Задачи статистического обучения

Набор методов для *изучения и понимания* данных:

- Методы обучения с учителем

Задачи статистического обучения

Набор методов для *изучения и понимания* данных:

- Методы обучения с учителем
- Методы обучения без учителя

Обучение с учителем

Имеется:

- Измерение Y (оно же: зависимая переменная, отклик, целевая переменная и т.п.)

Обучение с учителем

Имеется:

- Измерение Y (оно же: зависимая переменная, отклик, целевая переменная и т.п.)
- Вектор из p условий эксперимента X (вход, регрессоры, независимые переменные, ковариаты, признаки, ...)

Обучение с учителем

Имеется:

- Измерение Y (оно же: зависимая переменная, отклик, целевая переменная и т.п.)
- Вектор из p условий эксперимента X (вход, регрессоры, независимые переменные, ковариаты, признаки, ...)
- В задачах *регресии* Y — число (например, цена, давление крови, курс акций и т.п.)

Обучение с учителем

Имеется:

- Измерение Y (оно же: зависимая переменная, отклик, целевая переменная и т.п.)
- Вектор из p условий эксперимента X (вход, регрессоры, независимые переменные, ковариаты, признаки, ...)
- В задачах *регресии* Y — число (например, цена, давление крови, курс акций и т.п.)
- В задачах *классификации* Y — категориальная переменная, принимает конечное число значений из неупорядоченного множества (выжил/умер, цифра 0-9, тип заболевания)

Обучение с учителем

Имеется:

- Измерение Y (оно же: зависимая переменная, отклик, целевая переменная и т.п.)
- Вектор из p условий эксперимента X (вход, регрессоры, независимые переменные, ковариаты, признаки, ...)
- В задачах *регресии* Y — число (например, цена, давление крови, курс акций и т.п.)
- В задачах *классификации* Y — категориальная переменная, принимает конечное число значений из неупорядоченного множества (выжил/умер, цифра 0-9, тип заболевания)
- Доступна тренировочная выборка: $(x_1, y_1), \ldots, (x_N, y_N)$

Задачи

На основе тренировочной выборки мы хотим:

- Предсказать значения Y по X как на тренировочной выборке, так и для новых наблюдений
- Осознать, какие признаки влияют на отклик, и каким образом
- Оценить «качество» наших предсказаний и выводов

Немного философии

- Важно понять идеи, стоящии за различными методами, чтобы было понятно, какой метод в какой ситуации использовать
- Мы начнем с более простых методов, чтобы было проще понять более сложные
- Важно очень аккуратно оценить качество работы метода, чтобы понять, насколько хорошо он работает (часто простые методы работают почти так же хорошо, как и сложные!)

Обучение без учителя

- Нет зависимой переменной, только набор наблюдений

Обучение без учителя

- Нет зависимой переменной, только набор наблюдений
- Цель более «нечеткая»: найти группы наблюдений ведущих себя похоже, найти похожие признаки, найти комбинации признаков с большим разнообразием и т.п.

Обучение без учителя

- Нет зависимой переменной, только набор наблюдений
- Цель более «нечеткая»: найти группы наблюдений ведущих себя похоже, найти похожие признаки, найти комбинации признаков с большим разнообразием и т.п.
- Сложно понять, насколько все хорошо работает

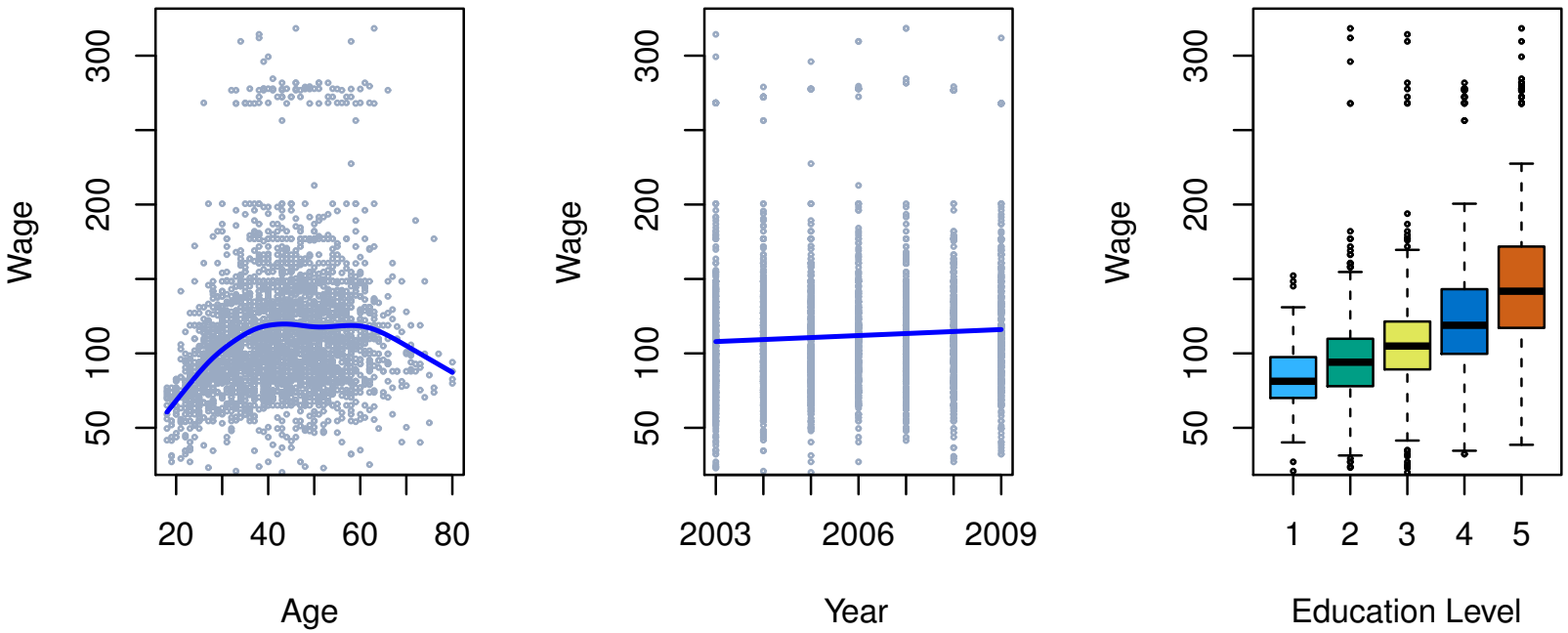
Обучение без учителя

- Нет зависимой переменной, только набор наблюдений
- Цель более «нечеткая»: найти группы наблюдений ведущих себя похоже, найти похожие признаки, найти комбинации признаков с большим разнообразием и т.п.
- Сложно понять, насколько все хорошо работает
- Отличается от обучения с учителем в подходах, часто используется как предварительный шаг

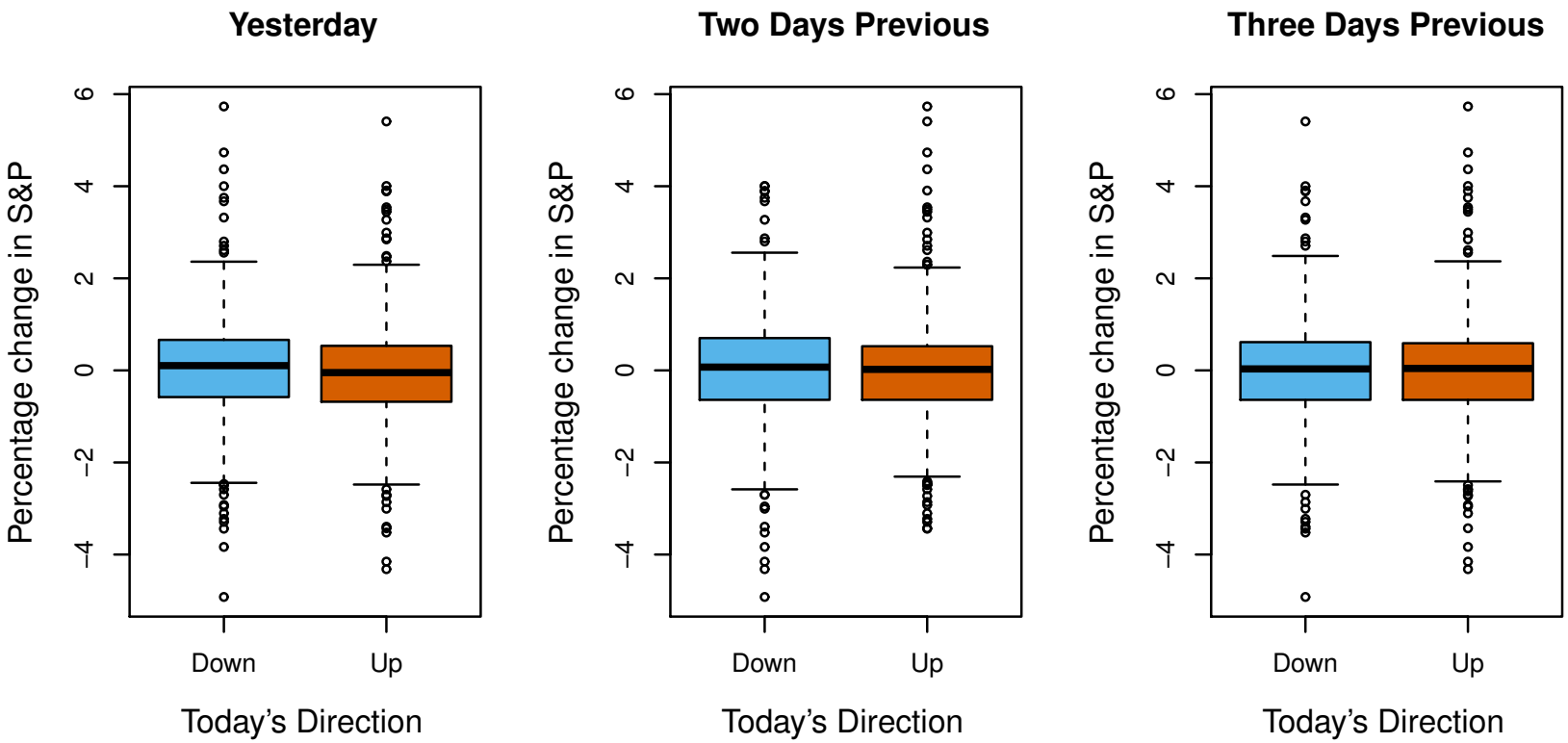
Почему не Machine Learning?

- Машинное обучение отпочковалось от методов искусственного интеллекта
- Статистическое обучение — раздел статистики
- У ML и SL много общего:
  - Машинное обучение концентрируется на анализе больших массивов данных и точности предсказания
  - Статистическое обучение занимается построением моделей, возможностью их интерпретации, применимости и т.п.
- В настоящее время отличия все более и более размываются
- Machine Learning в последнее время становится ругательством, используемым направо и налево

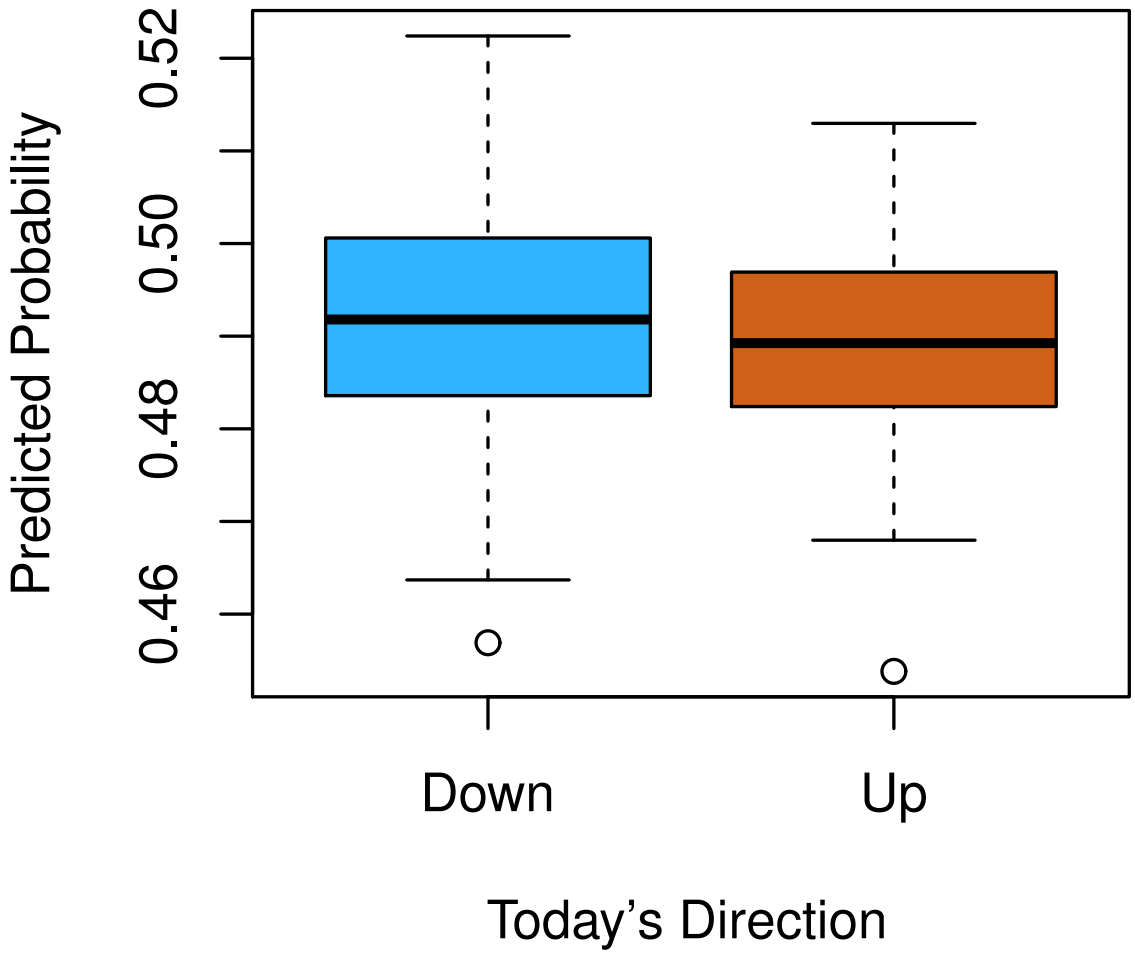
Пример: Wage


Wage
50
100
200
300
20
40
60
80
Age
Wage
50
100
200
300
2003
2006
2009
Year
Wage
50
100
200
300
1
2
3
4
5
Education Level

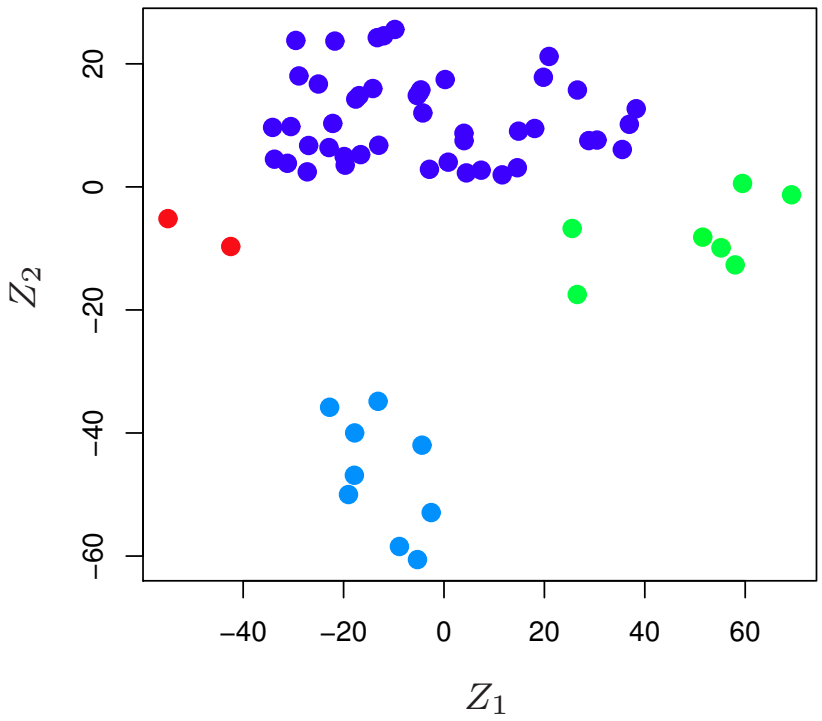
Пример: Stock

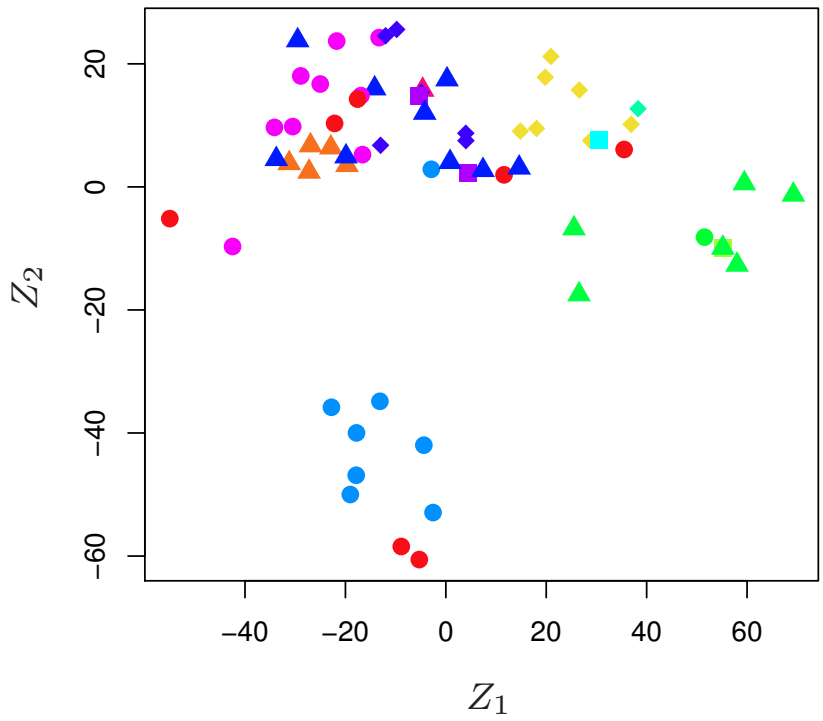


Пример: Stock



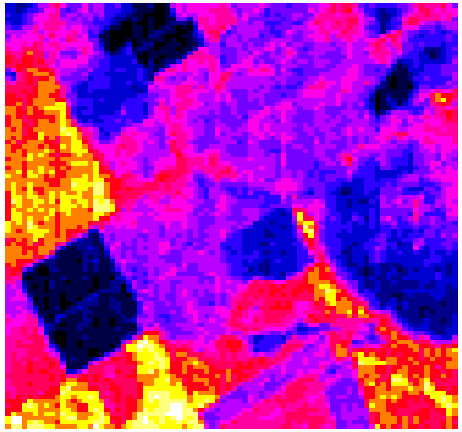
Пример: уровень экспрессии генов



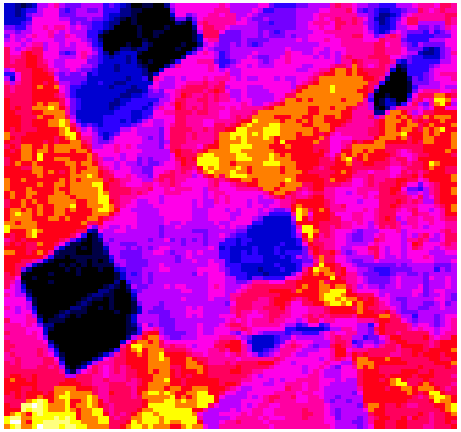


Пример: сегментация по типу использования почвы

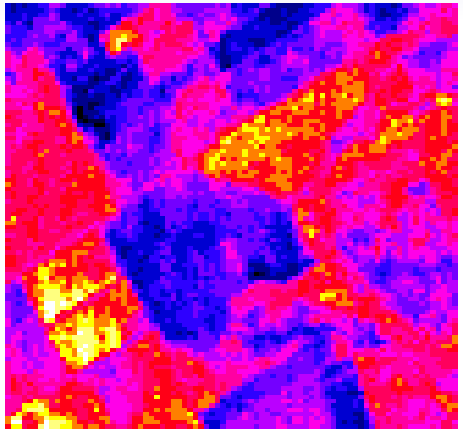
Spectral Band 1



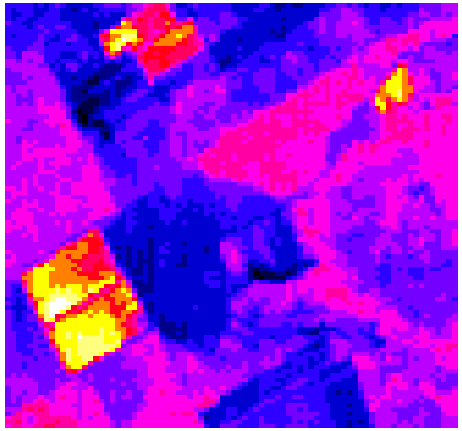
Spectral Band 2



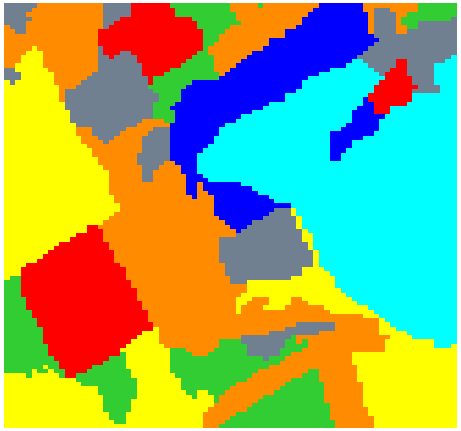
Spectral Band 3



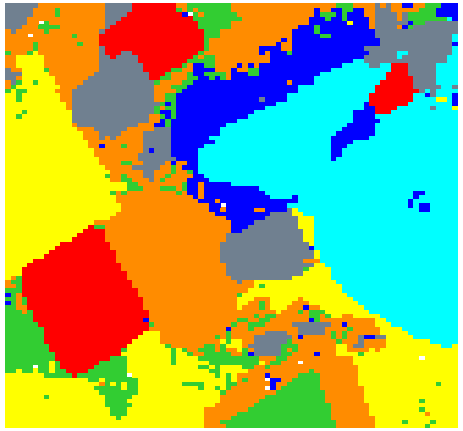
Spectral Band 4



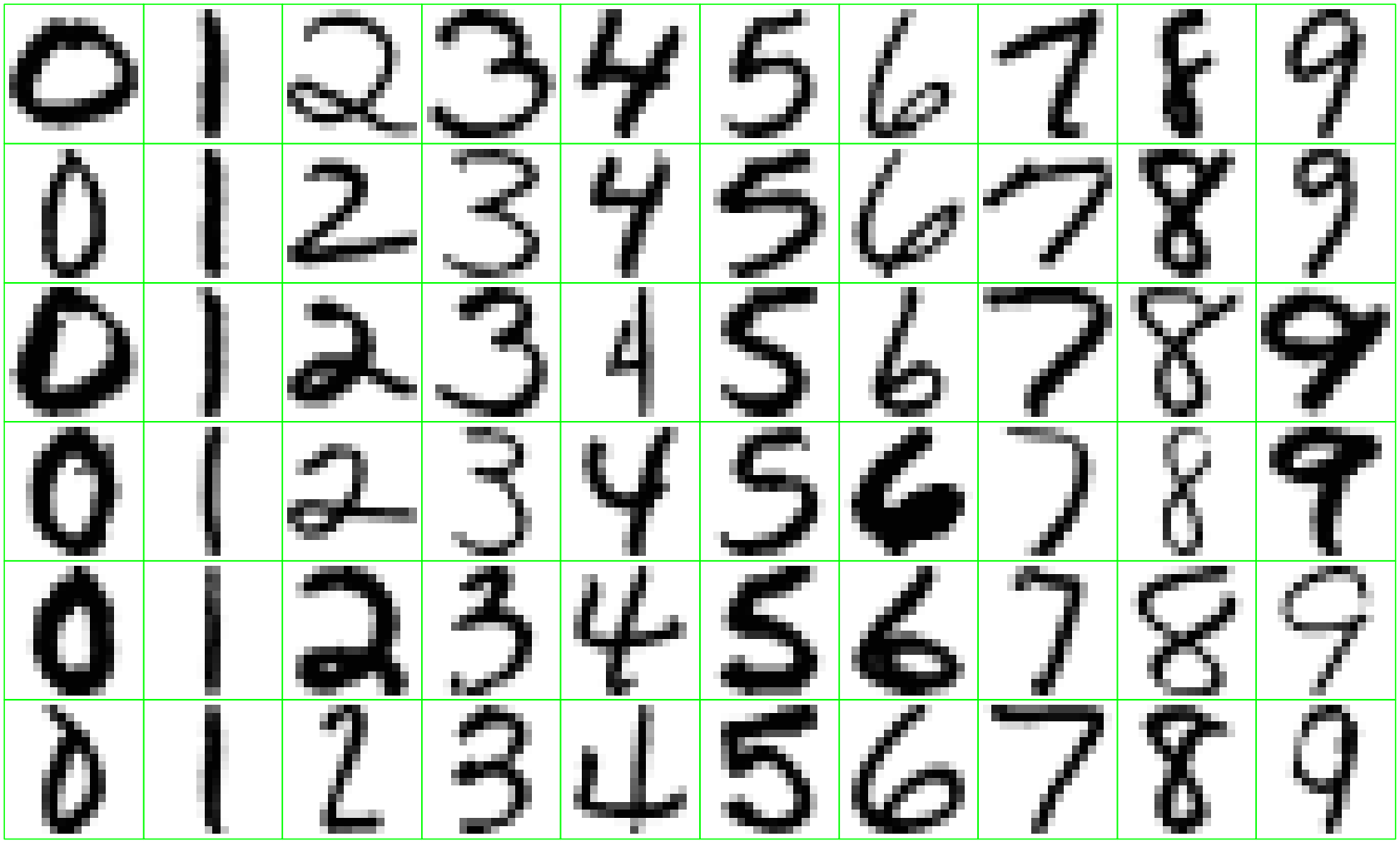
Land Usage



Predicted Land Usage



Пример: распознавание цифр



Литература

ISLR An Introduction to Statistical Learning with Applications in R. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.
`http://www-bcf.usc.edu/~gareth/ISL/`

ESL The Elements of Statistical Learning. Trevor Hastie, Robert Tibshirani, Jerome Friedman.
`http://statweb.stanford.edu/~tibs/ElemStatLearn/`