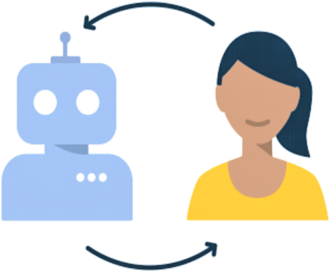
Активное обучение

Виктор Китов

victorkitov.github.io



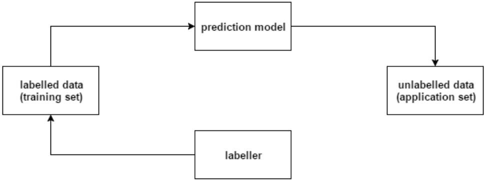
Курс поддержан фондом 'Интеллект'

Победитель конкурса VK среди курсов по IT

Содержание

Типовая последовательность действий

- Последовательность действий в обучении с учителем:



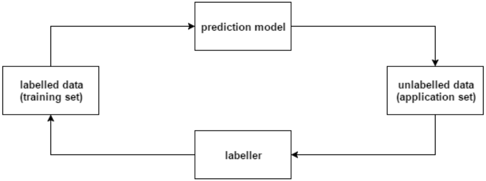
- Проблемы с точностью могут быть не из-за модели, а из-за обучающей выборки.
    - часто мы можем управлять созданием этой выборки.

Мотивация активного обучения

- Обычно доступно много неразмеченных данных:
  - документы, изображения, видео, речь в интернете
- Получить объекты легко, но дорого размечать.
  - хотим небольшую но высокоинформативную выборку
- Активное обучение (active learning, AL) - процесс оптимального построения обучающей выборки
  - за минимальное число разметок лучше всего настроить модель.

Схема активного обучения

- В активном обучении возвращаемся к расширению обучающей выборки много раз.



Мотивационный пример

- $x \in [0, 1]$, $y = \text{sign}(x - \theta)$
- Хотим оценить θ, используя N оценок $y(x)$.

[1] Какая будет точность, если объекты выбирать случайно из равномерного распределения?

Мотивационный пример

- $x \in [0,1]$, $y = \text{sign}(x - \theta)$
- Хотим оценить θ, используя N оценок $y(x)$.
- Объекты выбираем равномерно по сетке:[1]:
  - точность $O\left(\frac{1}{N}\right)$.

[1] Какая будет точность, если объекты выбирать случайно из равномерного распределения?

Мотивационный пример

- $x \in [0,1]$, $y = \text{sign}(x - \theta)$
- Хотим оценить θ, используя N оценок $y(x)$.
- Объекты выбираем равномерно по сетке:[1]:
  - точность $O\left(\frac{1}{N}\right)$.
- Выбираем каждый раз x в середине интервала $[x_i, x_{i+1}]$, где

$$i = \arg\max_k y(x_k) = -1$$

$$i + 1 = \arg\min_k y(x_k) = +1$$

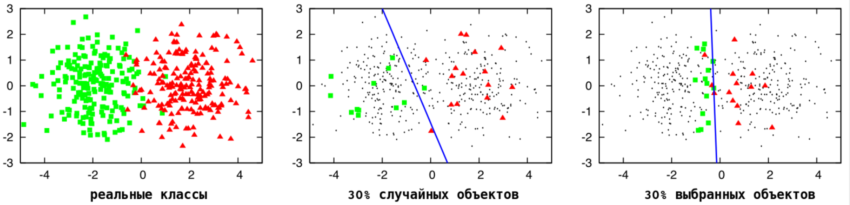
  - точность $O\left(\frac{1}{2^N}\right)$

[1] Какая будет точность, если объекты выбирать случайно из равномерного распределения?

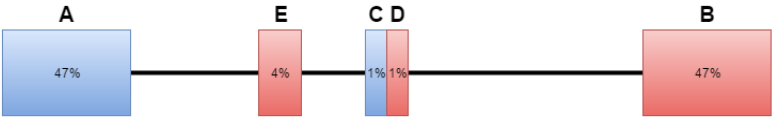
Мотивационный пример

$$x|y=-1\sim\mathcal{N}(\mu_1,\Sigma_1),\quad x|y=+1\sim\mathcal{N}(\mu_2,\Sigma_2)$$

Точность модели значительно повышается, если выбирать объекты вблизи разделяющей гиперплоскости.



Когда активное обучение хуже



- Предположим, объекты исходной выборки оказались из блоков A,C,D,B.
- Активное обучение будет уточнять границу между C,D
    - точность 4%
- Но лучшее решение - граница между A,E
    - точность 1%

Решение проблемы

- Чтобы такого не возникало,
нужна представительная начальная обучающая выборка.
- Также можно смешивать обычное и активное обучение (ε-**active стратегия**):

$$(x, y) \text{ выбирается } \begin{cases} \text{случайно} & \text{с вероятностью } \varepsilon \\ \text{направленно} & \text{с вероятностью } 1-\varepsilon \end{cases}$$

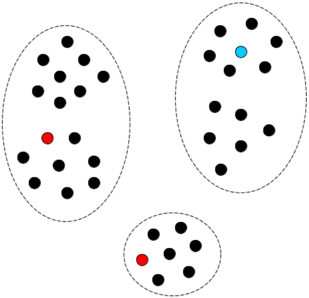
  - $\varepsilon \in [0, 1]$ контролирует exploration-exploitation tradeoff.
  - можно динамически менять ε (exponential gradient):
- Валидационная выборка AL - нерепрезентативна (сэмплируются более сложные объекты)
  - оценивать качество нужно на
случайной валидационной выборке.

Категоризация постановок задач и методов

- Объекты для разметки могут:
  - выбираться из готовой коллекции (в лекции)
  - выбираться из динамического потока
    - экспертная интерпретация изменения цены на отдельные акции
  - генерироваться вручную
    - в робототехнике можем посылать любое управляющее воздействие и смотреть результат.
- Основные подходы активного обучения
  - основанные на кластеризации
  - основанные на предиктивной модели

Содержание

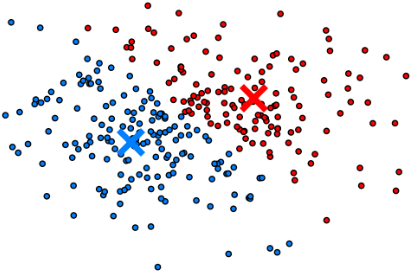
Кластерный подход к активному обучению



- Идея: более репрезентативны и разнообразны объекты из разных кластеров.
    - выбираем объекты стратифицированно по кластерам
- Результирующая выборка не привязана к целевой модели
    - поэтому результат может быть хуже
    - зато - применимость к разным моделям

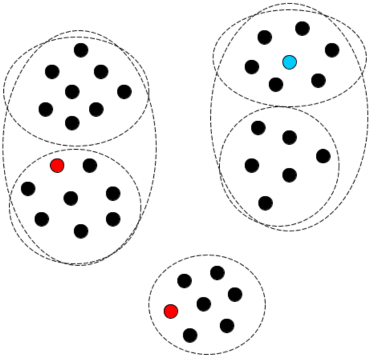
Проблемы кластерного подхода

Кластерная структура может отсутствовать:



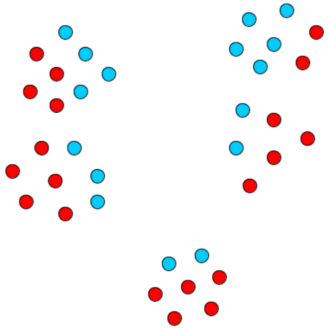
Проблемы кластерного подхода

Кластеры могут быть разного уровня детализации:



Проблемы кластерного подхода

Отклики могут быть различные внутри кластеров:



Содержание

Выбор по степени неуверенности: 2 класса

- **Сэмплирование по минимальной уверенности (least confident sampling):**
  - выбираем x, в прогнозе которого больше всего не уверены
- Бинарная классификация:

$$x^* = \arg\min_x |P(y = +1|x) - 0.5|$$

- Линейный классификатор без вероятностей:

Выбор по степени неуверенности: 2 класса

- **Сэмплирование по минимальной уверенности (least confident sampling):**
  - выбираем x, в прогнозе которого больше всего не уверены
- Бинарная классификация:

$$x^* = \arg\min_x \left|P(y=+1|x) - 0.5\right|$$

- Линейный классификатор без вероятностей:

$$x^* = \arg\min_x \frac{\left|w_0 + w^T x\right|}{\|w\|} = \arg\min_x \left|w_0 + w^T x\right|$$

Выбор по степени неуверенности: C классов

- **Сэмплирование по минимальной уверенности (least confident sampling)**[2]
  - объект с максимальной вероятностью ошибки:

$$x^* = \arg\max_x \left|1 - \max_y p(y|x)\right|$$

[2] Предложите модификацию этого метода, работающую для регрессии.

Выбор по степени неуверенности: C классов

- **Сэмплирование по минимальной уверенности (least confident sampling)**[2]
  - объект с максимальной вероятностью ошибки:

$$x^* = \arg\max_x \left|1 - \max_y p(y|x)\right|$$

- **Сэмплирование по отступу (margin sampling)**
  - объект с минимальным зазором в лидирующем классе:

$$x^* = \arg\min_x \{p(\widehat{y}_1|x) - p(\widehat{y}_2|x)\}$$

$$\widehat{y}_1 = \arg\max_y p(y|x), \quad \widehat{y}_2 = \arg\max_{y \neq \widehat{y}_1} p(y|x)$$

[2] Предложите модификацию этого метода, работающую для регрессии.

Выбор по степени неуверенности: C классов

- **Сэмплирование по минимальной уверенности (least confident sampling)**[2]
  - объект с максимальной вероятностью ошибки:

$$x^* = \arg\max_x \left| 1 - \max_y p(y|x) \right|$$

- **Сэмплирование по отступу (margin sampling)**
  - объект с минимальным зазором в лидирующем классе:

$$x^* = \arg\min_x \{p(\widehat{y}_1|x) - p(\widehat{y}_2|x)\}$$

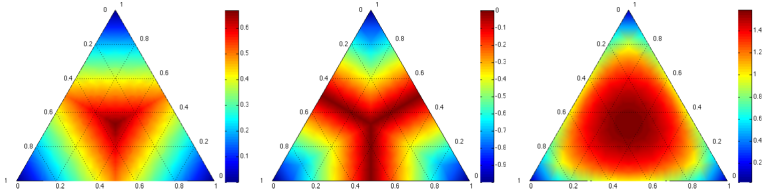
$$\widehat{y}_1 = \arg\max_y p(y|x), \quad \widehat{y}_2 = \arg\max_{y \neq \widehat{y}_1} p(y|x)$$

- **Сэмплирование по энтропии (entropy sampling)**
  - объект с максимальной энтропией класса:

$$x^* = \arg\max_x \{Entropy(y|x)\}, \quad Entropy(y|x) = -\sum_y p(y|x) \ln p(y|x)$$

[2] Предложите модификацию этого метода, работающую для регрессии.

Приоритетность от $p(y=1|x), p(y=2|x), p(y=3|x)$



least confident margin entropy

Отбор по несогласию в ансамбле моделей

- Классификация ансамблем моделей $f_1, ...f_M$; выбираем x, для которого $f_1(x), ...f_M(x)$ сильнее всего рассогласованы.
- **Энтропия голосования (vote entropy)**[3]

$$x^* = \arg\max_x \{Entropy(y|x)\}$$

$$Entropy(y|x) = -\sum_{c=1}^{C} \frac{\#[f_i(x) = c]}{M} \ln\left(\frac{\#[f_i(x) = c]}{M}\right)$$

[3] Предложите модификацию этого метода, работающую для регрессии.

Отбор по несогласию в ансамбле моделей

- Классификация ансамблем моделей $f_1, ...f_M$; выбираем x, для которого $f_1(x), ...f_M(x)$ сильнее всего рассогласованы.
- **Энтропия голосования (vote entropy)**[3]

$$x^* = \arg\max_x \{Entropy(y|x)\}$$

$$Entropy(y|x) = -\sum_{c=1}^{C} \frac{\#[f_i(x) = c]}{M} \ln\left(\frac{\#[f_i(x) = c]}{M}\right)$$

- **Рассогласованность распределений (spread in probabilities)**

$$x^* = \arg\max_x \frac{1}{M}\sum_{m=1}^{M} \rho(P_m(x), P(x)) \qquad P(x) = \frac{1}{M}\sum_{m=1}^{M} P_m(x)$$

$P_m(x)$ – распределение классов согласно f_m, ρ-напр. KL

[3] Предложите модификацию этого метода, работающую для регрессии.

Максимизация ожидаемого влияния на модель

- **Максимизация ожидаемого влияния на модель (expected model change)**
  - выберем x сильнее всего изменяющий модель.

[4] Важна нормализация признаков, т.к. их масштаб влияет на $\nabla_\theta \mathcal{L}(f_\theta(x), y)$.

Максимизация ожидаемого влияния на модель

- **Максимизация ожидаемого влияния на модель (expected model change)**
  - выберем x сильнее всего изменяющий модель.
- После добавления (x, y):

$$\nabla L(\theta) = \underbrace{\sum_{n=1}^{N} \nabla \mathcal{L}(f_\theta(x_n), y_n)}_{\approx 0 \text{ т.к. оптимизация сошлась}} + \nabla \mathcal{L}(f_\theta(x), y) \approx \nabla \mathcal{L}(f_\theta(x), y)$$

$$\|\nabla_\theta \mathcal{L}(f_\theta(x), y)\| \approx \text{ влияние } (x, y) \text{ на модель}$$

[4]Важна нормализация признаков, т.к. их масштаб влияет на $\nabla_\theta \mathcal{L}(f_\theta(x), y)$.

Максимизация ожидаемого влияния на модель

- **Максимизация ожидаемого влияния на модель (expected model change)**
  - выберем x сильнее всего изменяющий модель.
- После добавления (x, y):

$$\nabla L(\theta) = \underbrace{\sum_{n=1}^{N} \nabla \mathcal{L}(f_\theta(x_n), y_n)}_{\approx 0 \text{ т.к. оптимизация сошлась}} + \nabla \mathcal{L}(f_\theta(x), y) \approx \nabla \mathcal{L}(f_\theta(x), y)$$

$$\|\nabla_\theta \mathcal{L}(f_\theta(x), y)\| \approx \text{ влияние } (x, y) \text{ на модель}$$

- Выберем x максимизирующий ожидаемое изменение[4]:

$$x^* = \arg\max_x \sum_y p(y|x) \|\nabla_\theta \mathcal{L}(f_\theta(x), y)\|$$

[4] Важна нормализация признаков, т.к. их масштаб влияет на $\nabla_\theta \mathcal{L}(f_\theta(x), y)$.

Ожидаемое сокращение ошибки

- Предыдущие методы основывались на неопределенности отдельного объекта.
- **Ожидаемое сокращение ошибки (expected error reduction)**
  - выберем x, максимально снижающий неопределенность классов в неразмеченной выборке
  - наилучший эффект, но самый медленный метод
- Обозначим U - множество неразмеченных объектов.
- $f^{+(x,y)}$: модель f, дообученная на (x, y)

Ожидаемое сокращение ошибки

- **Снижение частоты ошибок (certainty improvement)**:

$$x^* = \arg\max_{x \in U} \sum_y p_f(y|x) \left(\sum_{u \in U} p_{f+(x,y)} \left(\widehat{y}_u | x_u \right) \right)$$

$$\widehat{y}_u = \arg\max_y p_{f+(x,y)}(y|x_u)$$

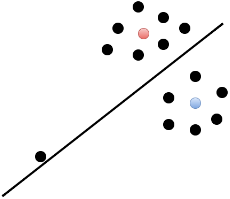
- **Снижение энтропии классов (entropy minimization)**:
  - эквивалентно максимизации логарифма правдоподобия

$$x^* = \arg\min_{x \in U} \sum_y p_f(y|x) \sum_{u \in U} \underbrace{-\sum_{\tilde{y}} p_{f+(x,y)}(\tilde{y}|x_u) \ln p_{f+(x,y)}(\tilde{y}|x_u)}_{Entropy(\tilde{y}|x_u,\, f^{+(x,y)})}$$

Обработка выбросов

Выбросы по умолчанию предпочитаются методами AL:

- лежат в отдельных "кластерах"
- модели ансамбля экстраполируют зависимость по-разному
- сильно влияют на модель и ее прогнозы



Учет выбросов

- Решение - отфильтровать выбросы либо занижать их влияние:

$$x^* = \arg\max_x \left\{ score(x) \times typicalness(x)^\beta \right\}$$

$$typicalness(x)^\beta \in [0,1] \text{ - типичность объекта}$$

$$\beta \geq 0 \text{ - гиперпараметр (сила фильтрации выбросов)}$$

- Примеры:

$$typicalness(x) = p(x)$$

$$typicalness(x) = \frac{1}{\rho\left(x, x_{i(K)}\right)}$$

$x_{i(K)}$- K-й ближайший сосед

Заключение

- Задача активного обучения: $\downarrow$размер и $\uparrow$информативность обучающей выборки.
  - обучающая выборка: баланс между случайной и направленной генерацией
  - валидационная выборка: только случайная.
- Подходы активного обучения:
  - основанные на кластеризации
  - основанные на модели:
    - по степени неуверенности прогнозов
    - по рассогласованости базовых моделями ансамбля
    - по влиянию на модель
    - по однозначности разметки тестовых объектов
- Выбросы должны отбрасываться или учитываться с меньшим весом.