



ТЕХНОСФЕРА

Лекция 5 Вопросы построения обучающих множеств

Владимир Гулин

5 октября 2018 г.

План лекции

Машинное обучение в реальной жизни

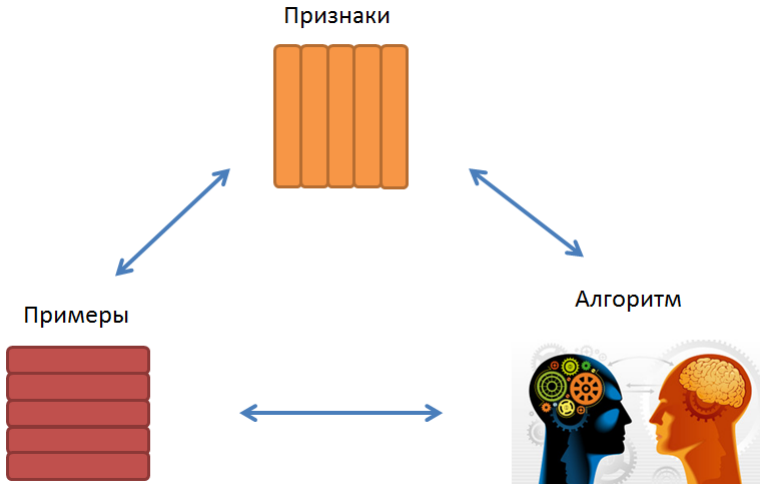
Sampling

Active Learning

Active Learning in practice

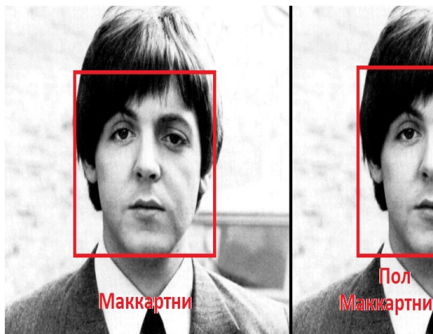
Машинное обучение в реальной жизни

Схема компонентов системы машинного обучения



Машинное обучение в реальной жизни

Ожидание

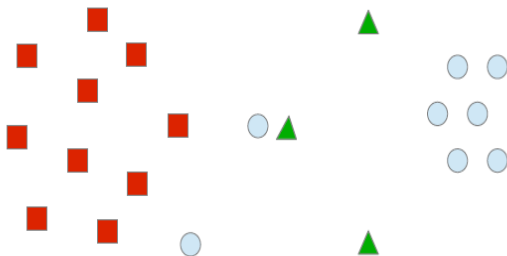


Реальность



- ▶ Работаем не на той же выборке, на которой обучались

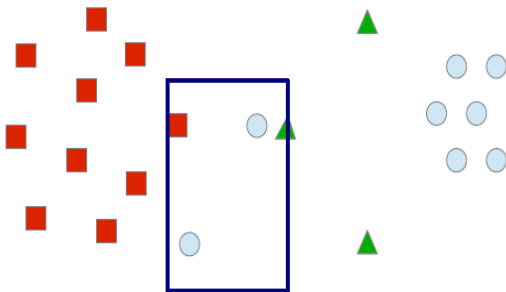
Классические проблемы обучающего множества



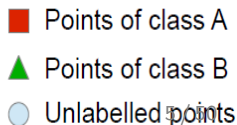
- Какие потенциальные проблемы с данными вы видите на этой картинке?

■ Points of class A
▲ Points of class B
○ Unlabelled points

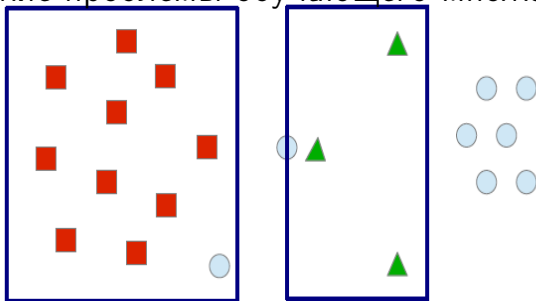
Классические проблемы обучающего множества



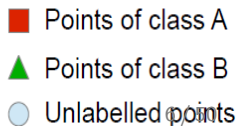
- Существуют неразмеченные точки на границе между классами



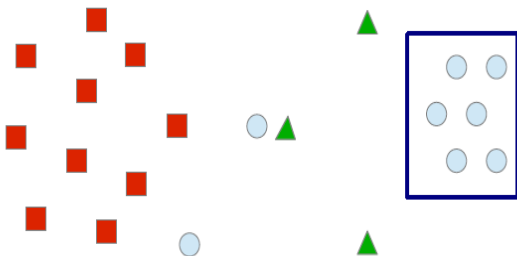
Классические проблемы обучающего множества



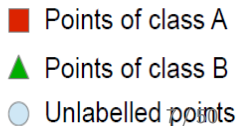
- ▶ Существуют неразмеченные точки на границе между классами
- ▶ Количество данных в разных классах несбалансировано



Классические проблемы обучающего множества



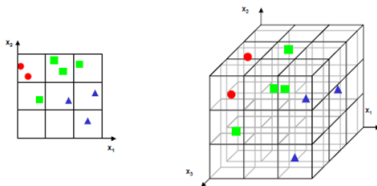
- ▶ Существуют неразмеченные точки на границе между классами
- ▶ Количество данных в разных классах несбалансировано
- ▶ Имеется неразмеченная группа данных



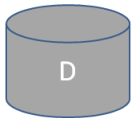
Несмещенная выборка

Определение

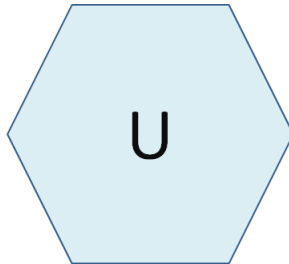
Несмещенная (репрезентативная) выборка - это такая выборка, в которой все основные признаки генеральной совокупности, из которой извлечена данная выборка, представлены приблизительно в той же пропорции или с той же частотой, с которой данный признак выступает в этой генеральной совокупности.



Labeled & Unlabeled data



Размеченное множество

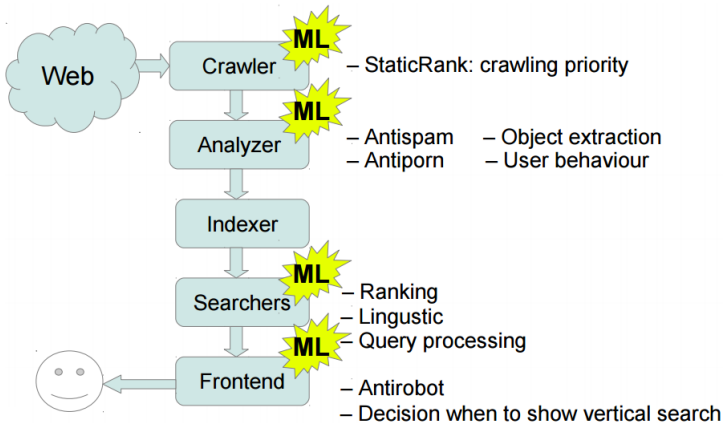


Неразмеченное множество

- ▶ Sampling
- ▶ Active Learning
- ▶ Semi-supervised learning

Мотивация

Упрощенная схема поисковой системы



Мотивация

Проблемы

- ▶ Все компоненты используют машинное обучение с учителем
- ▶ Ассесорские оценки дорогие удовольствие
- ▶ Требуются большие обучающие выборки для высокого качества
- ▶ Долго обучаться (примеров $10^6 - 10^7$)

Хотим компактные обучающие выборки

- ▶ Проще анализировать данные
- ▶ Быстрее можно перестраивать модели и проводить эксперименты

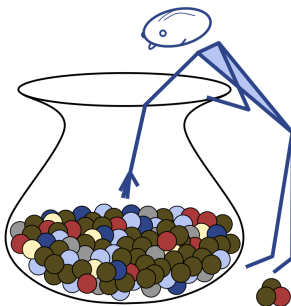
Sampling

Определение

Sampling - метод исследования множества, путем анализа его подмножеств

Применение

- ▶ Предварительный анализ данных
- ▶ Исходное множество слишком велико



Sampling

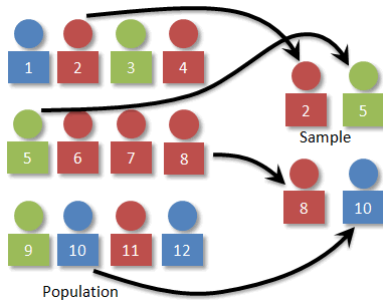
Simple random sampling

Systematic sampling

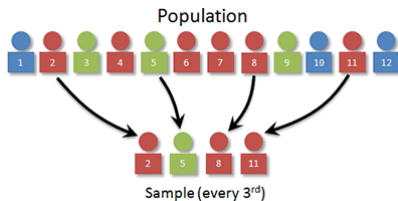
Stratified random sampling

Cluster sampling

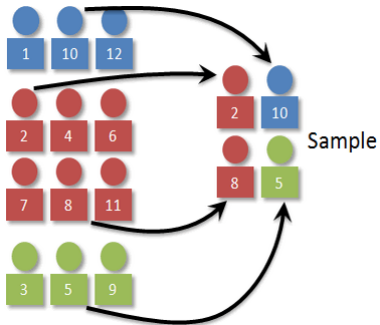
Simple random sampling



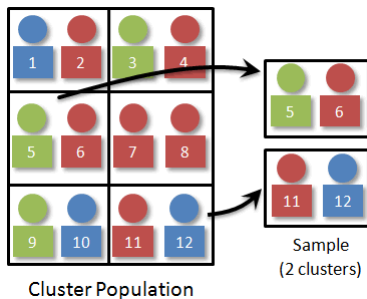
Systematic sampling



Stratified sampling



Cluster sampling

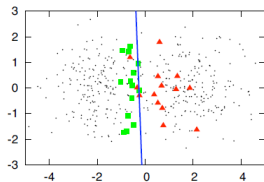
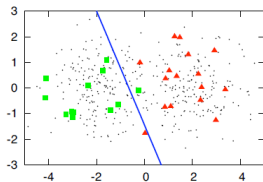
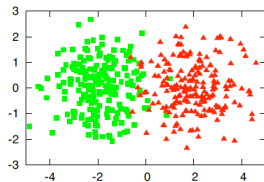


Вопрос:

А какой алгоритм семплирования выбрать?

Active Learning

Интуиция



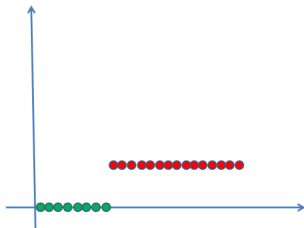
Active Learning

Идея

The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose that data from which it learns.

- ▶ Забываем на требование несмещенности выборки

Мотивирующий пример



$$g(x, \theta) = \begin{cases} 1 & x > \theta \\ 0 & \text{otherwise} \end{cases}$$

Вопрос:

Сколько точек необходимо для того, чтобы найти θ с точностью ε

Классическая схема машинного обучения

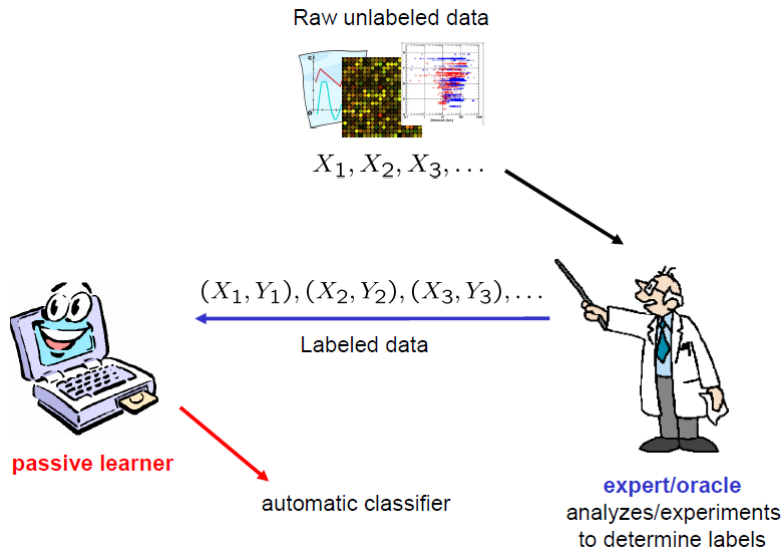
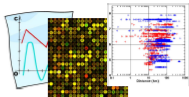


Схема с активным обучением

Raw unlabeled data



X_1, X_2, X_3, \dots

Learner requests labels
for **selected** data



active learner

$(X_1, ?)$

(X_1, Y_1)

$(X_3, ?)$

(X_3, Y_3)



expert/oracle

analyzes/experiments
to determine labels

automatic classifier

Типы активного обучения

Pool-based sampling

Stream based selective sampling

Query-synthezis

Active Learning Strategies

Uncertainty Sampling

Query-by-Committee

Expected Model Change

Expected Error Reduction

Variance Reduction

Density-Weighted Methods

Bonus technique

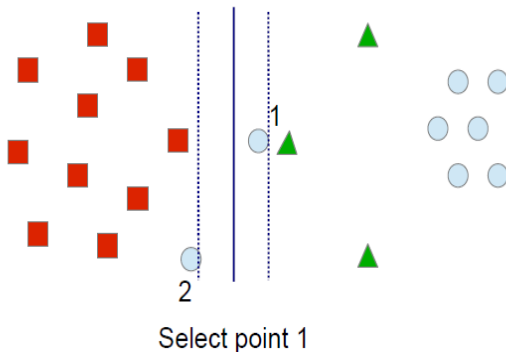
Идея

Построим модель на ошибках предыдущей и будем ей предсказывать точки, которые брать в обучение

$$L(y, h(\mathbf{x})) = \sum_{i=1}^N (y_i - h(\mathbf{x}_i))^2$$

$$\hat{f}(\mathbf{x}_i) = |y_i - h(\mathbf{x}_i)|, \quad i = 1, \dots, N$$

Uncertainty Sampling



Идея

Выбираем те примеры, в которых модель уверена меньше всего

$$x^* = \arg \min_x |P(\hat{y}|x) - 0.5|$$

- Points of class A
- ▲ Points of class B
- Unlabelled points

Uncertainty Sampling

Случай нескольких классов

- ▶ Least confident

$$x_{LC}^* = \arg \max_x (1 - P_\theta(\hat{y}|x))$$

$$\hat{y} = \arg \max_y P_\theta(y|x)$$

- ▶ Margin sampling

$$x_M^* = \arg \min_x P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)$$

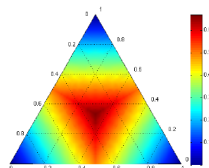
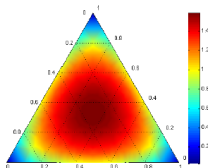
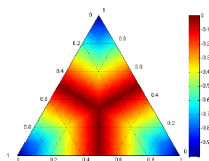
- ▶ Entropy (общий случай)

$$x_H^* = \arg \max_x - \sum_c P_\theta(\hat{y}_c|x) \log P_\theta(\hat{y}_c|x)$$

Uncertainty Sampling

Пример: Случай трехклассовой классификации

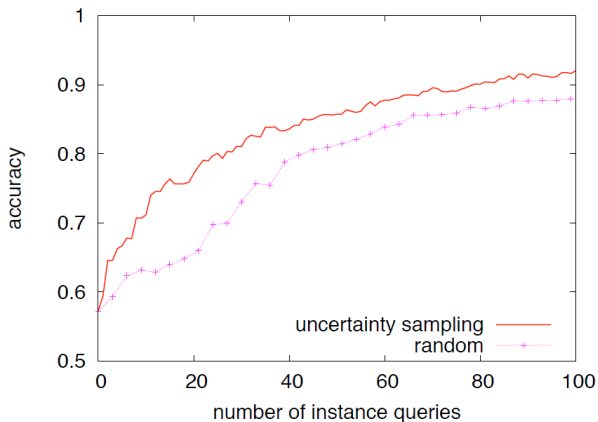
$$p_1 + p_2 + p_3 = 1$$



Вопрос:

- Какой мере неопределенности соответствует каждая из этих картинок?

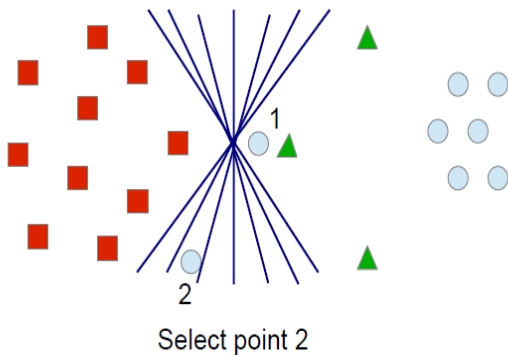
Uncertainty Sampling vs Random Sampling



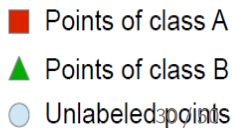
Вопрос:

- Что делать если у нас нет постериорного распределения $p(y|x)$?

Query-by-Committee



- Вместо одной модели используем коммитет



Query-by-Committee

Для измерения уровня несогласия между моделями используют:

Vote Entropy

$$x_{VE}^* = \arg \max_x - \sum_c \frac{V(y_c)}{T} \log \frac{V(y_c)}{T}$$

Kullback-Leibler Divergence

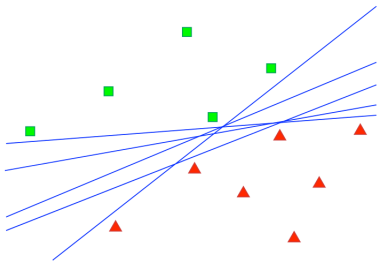
$$x_{KL}^* = \arg \max_x \frac{1}{T} \sum_{t=1}^T D(P_{\theta^t} || P_T), \quad \text{где}$$

$$D(P_{\theta^t} || P_T) = \sum_c P_{\theta^t}(y_c|x) \log \frac{P_{\theta^t}(y_c|x)}{P_T(y_c|x)}, \quad P_T(y_c|x) = \frac{1}{T} \sum_{t=1}^T P_{\theta^t}(y_c|x)$$

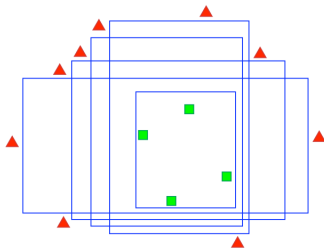
Query-by-Committee

Идея:

Выбираем очередную точку максимально сокращая пространство решений



(a)



(b)

Query-by-Bagging

Qbag

Input: T – initial labelled training set

C – size of the committee

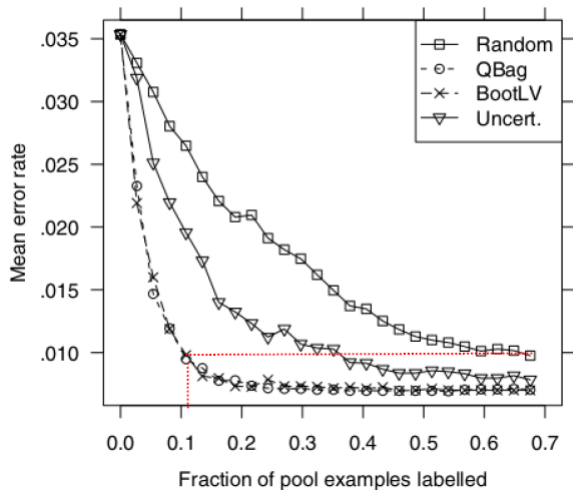
A – learning algorithm

U – set of unlabelled objects

Output: T' – extended training set

1. Uniformly resample T , obtain $T_1 \dots T_C$, where $|T_i| < |T|$
2. For each T_i build model M_i using A
3. Select $x^* = \min_{x \in U} | \sum_{i=1}^C I(M_i = 1) - \sum_{i=1}^C I(M_i = 0) |$
4. Pass x^* to assessor and update T
5. Repeat from 1 until convergence

Query-by-Bagging



Query-by-Boosting

Вспоминаем AdaBoost

1. Инициализировать веса объектов $w_j = 1/N, j = 1, 2, \dots, N$.
2. Для всех i от 1 до T :

(a) Построить классификатор $a_i(\mathbf{x})$, используя веса w_j

(b) Вычислить

$$err_i = \frac{\sum_{j=1}^N w_j I(y_j \neq a_i(\mathbf{x}_j))}{\sum_{j=1}^N w_j}$$

(c) Вычислить

$$b_i = \log \frac{1 - err_i}{err_i}$$

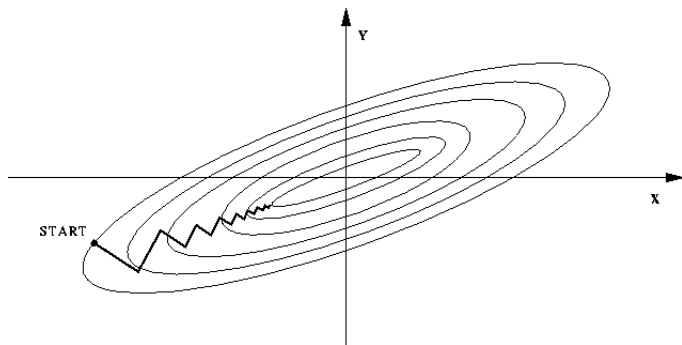
(d) Присвоить $w_j \rightarrow w_j \cdot \exp[b_i \cdot I(y_j \neq a_i(\mathbf{x}_j))], j = 1, \dots, N$.

(e) Нормируем веса объектов

$$w_j \rightarrow \frac{w_j}{\sum_{j=1}^N w_j}, j = 1, \dots, N.$$

3. $h(\mathbf{x}) = \text{sign} \left[\sum_{i=1}^T b_i a_i(\mathbf{x}) \right]$

Expected Model Change



Идея

Выбираем примеры, оказывающие наибольшее влияние на модель

$$x_{EMC}^* = \arg \max_x \sum_c P_\theta(y_c|x) \|\nabla L_\theta(D \cup (x, y_c))\|$$

При этом надо понимать

$$\|\nabla L_\theta(D \cup (x, y_c))\| \approx \|\nabla L_\theta(x, y_c)\|$$

Expected Error Reduction

Идея

Выбираем примеры, увеличивающие обобщающую способность нашей модели

Замечание

- ▶ Необходимо научиться оценивать ошибку обобщения модели на данных $D \cup (x, y)$
- ▶ В качестве валидационной выборки будем использовать все оставшееся неразмеченное множество U

$$x_{0/1}^* = \arg \min_x \sum_c P_\theta(y_c|x) \left(\sum_{u=1}^U 1 - P_{\theta+(x,y_c)}(\hat{y}|x^{(u)}) \right)$$

$$x_{log}^* = \arg \min_x \sum_c P_\theta(y_c|x) \left(- \sum_{u=1}^U \sum_k P_{\theta+(x,y_c)}(y_k|x^{(u)}) \log P_{\theta+(x,y_c)}(y_k|x^{(u)}) \right)$$

Variance Reduction

Идея

Можем минимизировать ошибку обобщения неявно, уменьшая variance модели

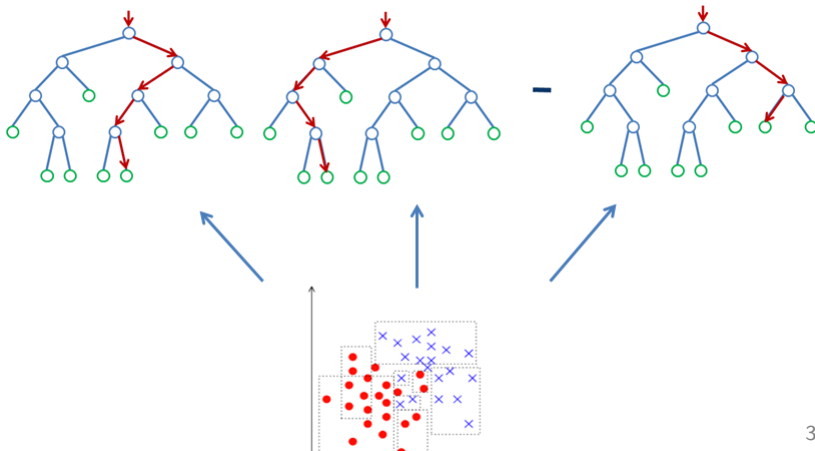
$$E[(\hat{y} - y)^2|x] = \textit{Noise} + \textit{Bias}^2 + \textit{Variance}$$

- Каким образом уменьшать Variance?

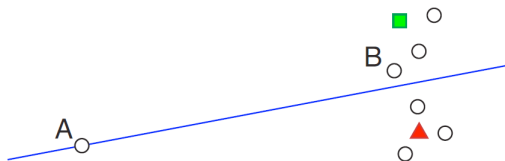
Variance Reduction

Идея

Будем собирать примеры, которые попадают в листы нашей модели, соответствующие малому числу примеров обучающей выборки



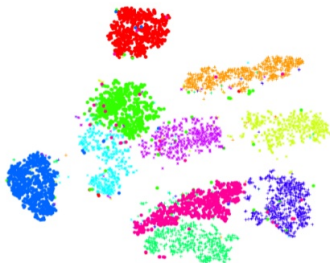
Density-Weighted Methods



Идея

Будем дополнительно использовать информацию о схожести примеров при добавлении новых, чтобы отбираемые примеры были “репрезентативны” относительно данного распределения

Density-Weighted Methods



$$x^* = \arg \max_x \phi_A(x) \times \left(\frac{1}{U} \sum_{u=1}^U \rho(x, x^u) \right)^\beta$$

ϵ -active

Идея

Будем с некоторой вероятностью смотреть и в другие области пространства

Algorithm 1 ϵ -active

- 1: **Input:** X, ϵ
 - 2: **Output:** x_t, r_t
 - 3: $x_t = \begin{cases} \text{Activelearning}(X) & \text{if}(q < \epsilon) \\ \text{Random}(X) & \text{if}(q \geq \epsilon) \end{cases}$
 - 4: **if** x was not queried in the past **then** Query O for label y of x
 - 5: Observe reward r_t
-

EG-active

Exponentiated gradient active

Будем подбирать вероятность динамически

Algorithm 2 EG-active.

Input: $(\epsilon_1, \dots, \epsilon_T)$: candidate values for ϵ

β, τ and k : parameters for EG

N : number of iterations

$p_k \leftarrow \frac{1}{T}$ and $w_k \leftarrow 1, k = 1, \dots, T$

for $i=1$ **to** N **do**

 Sample d from discrete (p_1, \dots, p_T)

 Run the ϵ -active with ϵ_d

 Receive the feedback r_t

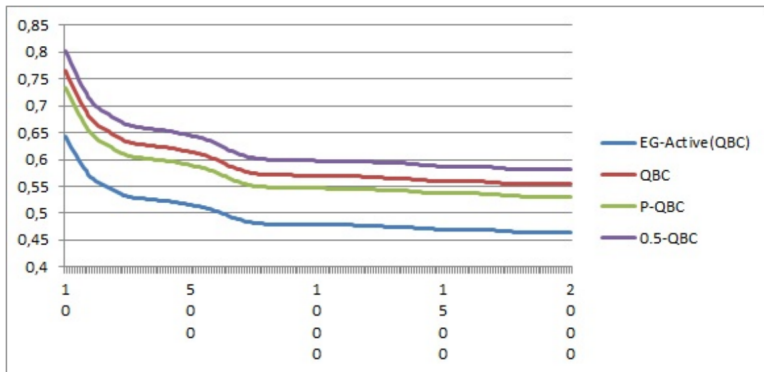
$w_k \leftarrow w_k \exp(\frac{\tau[r_i I(k=d) + \beta]}{p_k}), k = 1, \dots, T$

$p_k \leftarrow (1 - k)(\frac{w_k}{\sum_{j=1}^T w_j} + \frac{k}{T}), k = 1, \dots, T$

end for

EG-active

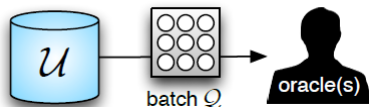
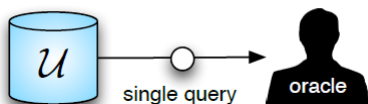
Результаты экспериментов на UCI



Active Learning in practice

Querying in Batches

Вместо того, чтобы давать экспертам примеры по одному, отдадим сразу пачку



Вопрос:

- ▶ Как правильно организовать эту процедуру?

Active Learning in practice

Noisy Oracles

- ▶ Эксперты и люди и они совершают ошибки
- ▶ Необходимо проверять оценки экспертов другими экспертами
- ▶ НЕ каждый эксперт знает правильный ответ при разметке (нужно эксперты в узких областях)
- ▶ группа экспертов \neq миллионы пользователей (смещенные оценки)

Labeling costs

- ▶ Экспертам надо платить зарплату
- ▶ Много экспертов \rightarrow много денег
- ▶ Что лучше? Уточнить оценку для уже известного примера или оценить новый?

Итоги

- ▶ Активное обучение простой эффективный метод для набора датасета
- ▶ Может быть применено практически для любых методов машинного обучения с учителем
- ▶ Требуется значительных вычислительных расходов
- ▶ Собранный датасет работает только для данных признаков и для данного алгоритма. Если, что-то меняется, то похорошему активно обучаться надо заново
- ▶ Тестовый датасет всегда должен быть репрезентативен!!!

Задача

Дано: Имеется набор точек из 10 мерного пространства данных.

Требуется: Требуется реализовать процедуру активного обучения для решения задачи регрессии.

Пошаговая инструкция

1. Скачать данные и запустить шаблон кода на python
<https://goo.gl/MDZNaX>

```
$ python al.py -h  
$ python al.py -tr train.txt -te test.txt
```

2. Выбрать алгоритм для решения задачи регрессии
3. Выполнить random sampling
4. Разработать процедуру активного обучения
5. Построить графики rmse, в зависимости от числа примеров

Дз по активному обучению:

Задание:

Реализовать один из алгоритмов активного обучения, рассказанных на лекции и применить его в соревновании на Kaggle.

Вопросы

