Машинное обучение: качество

ИГОРЬ КУРАЛЁНОК

К.Ф.-М.Н., ЯНДЕКС/СПБГУ

Немного «фраз»

«If you can't measure it, you can't improve it»

— Lord Kelvin

«Гораздо легче что-то измерить, чем понять, что именно вы измеряете.»

— Джон Уильям Салливан

Постановка в случае учителя

$$F_0 = \underset{F=A(X)}{\operatorname{argmin}}\ \mu\left(Loss(y, F(x))\right)$$

- Ожидание хотим считать по всей ген. совокупности
- Функцию обучаем на X

=> Если бы X была репрезентативной то все проще:

$$F_0 = \underset{F}{\operatorname{argmin}} \sum_{X} Loss\left(y_i, F(x_i)\right)$$

Какая нужна выборка

Иными словами, репрезентативная выборка представляет собой микрокосм, меньшую по размеру, но точную модель генеральной совокупности, которую она должна отражать.

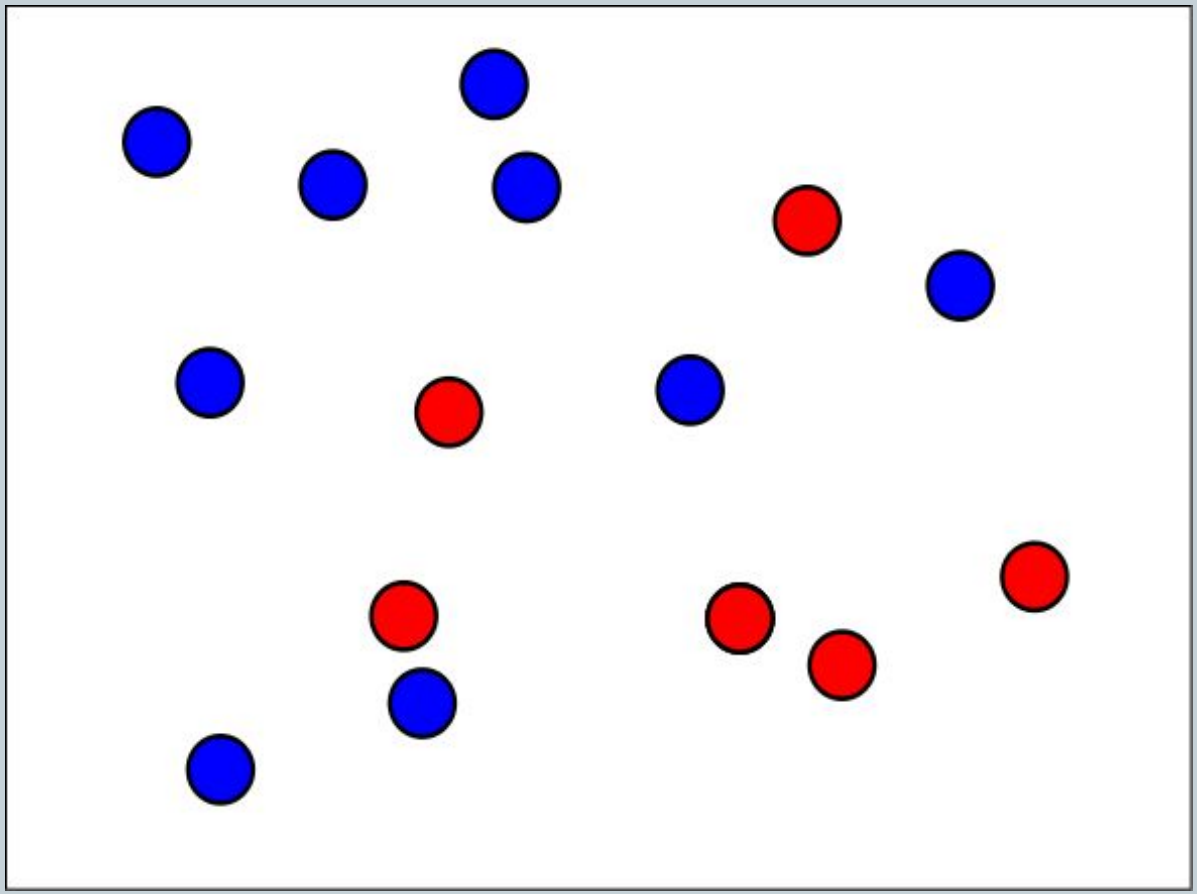
- Дж. Б. Мангейм, Р. К. Рич

Интересно получить выборку, несмещенную (смещенную не более чем ...) по результатам процедуры обучения:

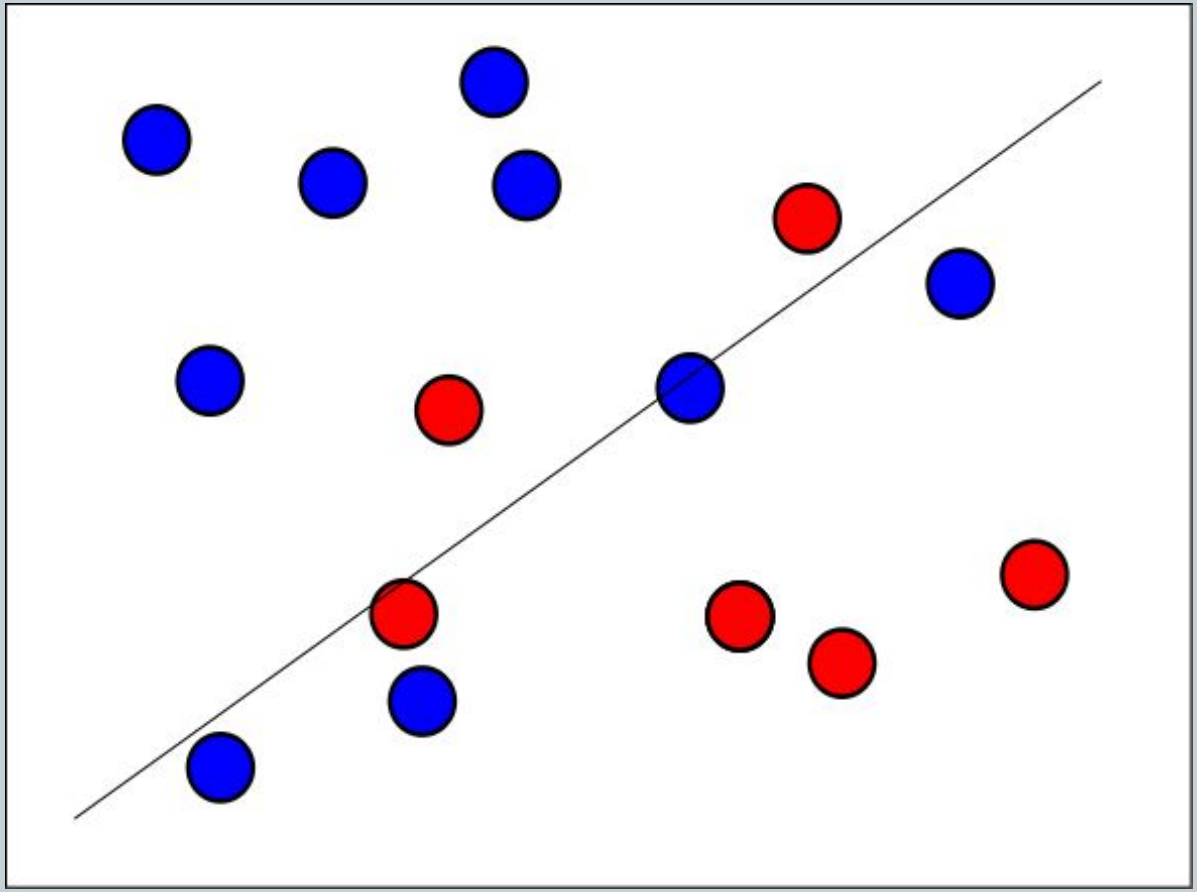
- Найти «хороший» способ генерации выборки при условии процедуры подбора
- Наложить ограничения на процедуру подбора
- Ограничения на решающую функцию

=> Надо научиться мерять смещенность выборки

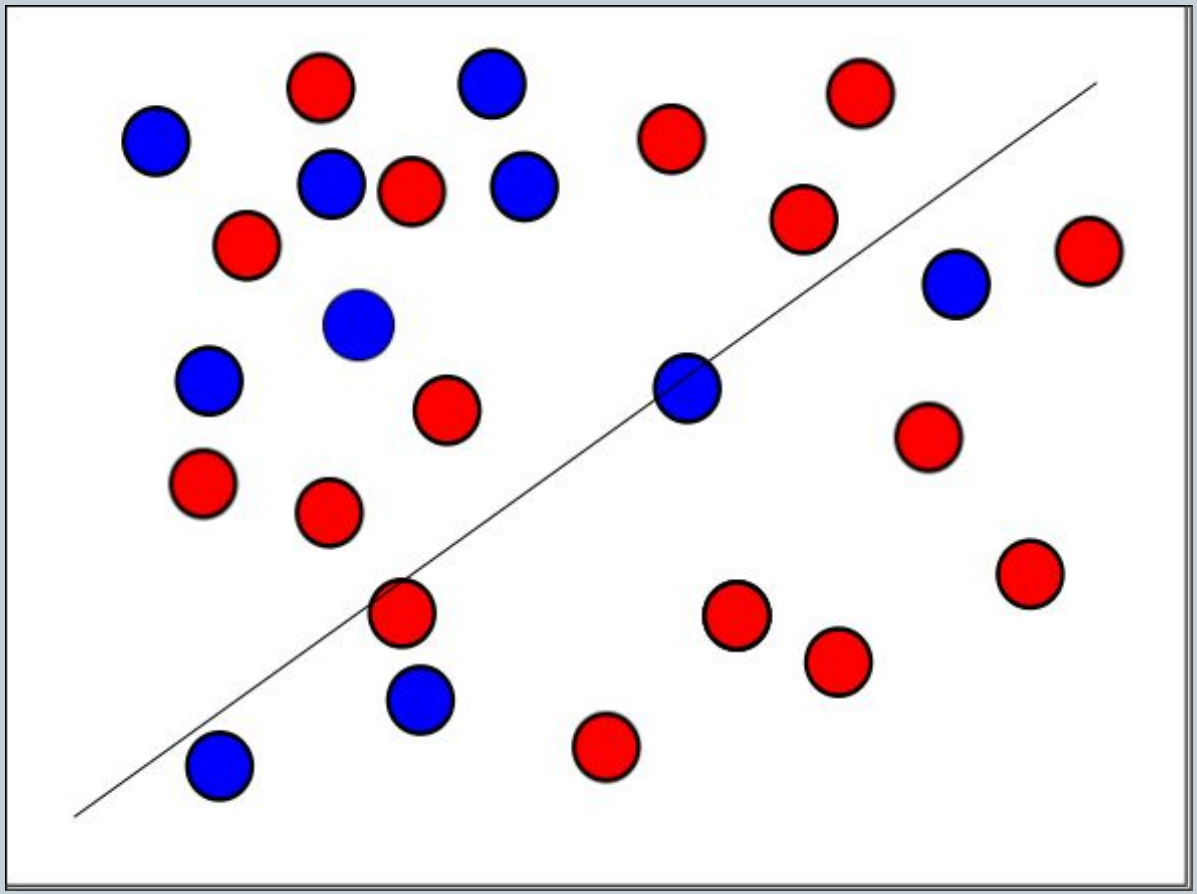
Как это выглядит на практике?



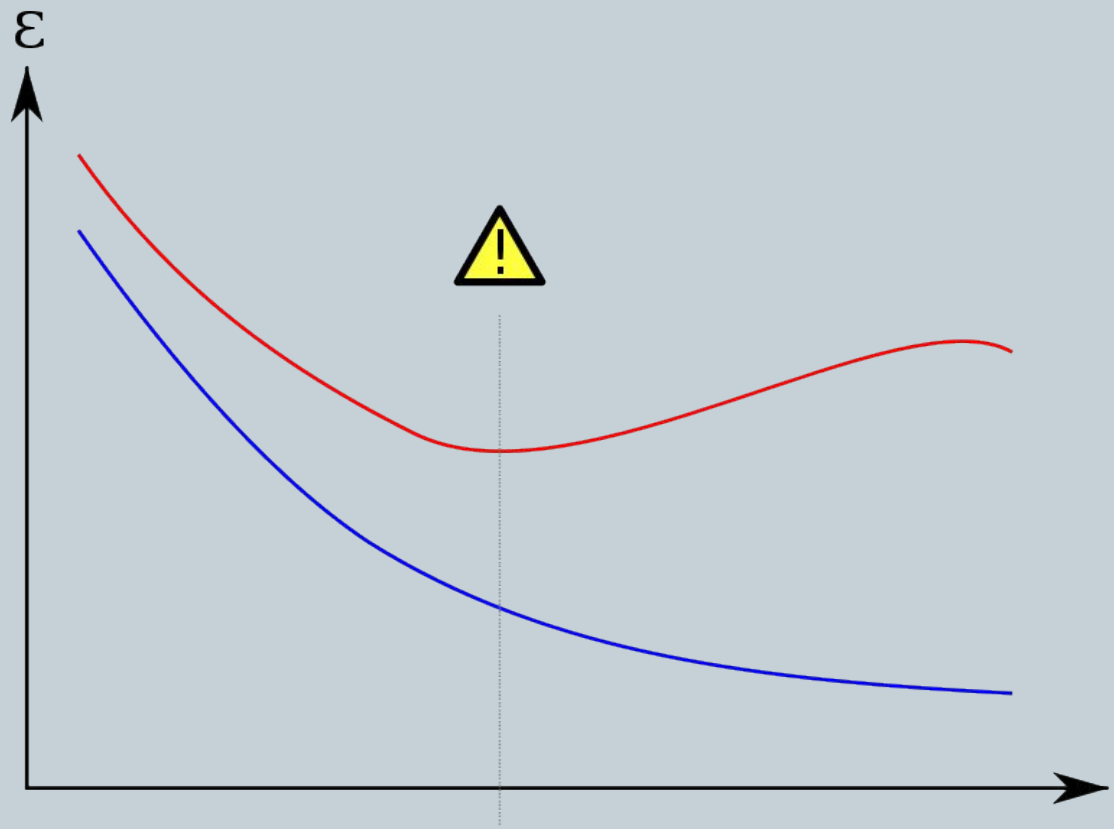
Как это выглядит на практике?



Как это выглядит на практике?



Виды ошибок обучения



Виды ошибок обучения

Переобучение, переподгонка (overtraining, overfitting) — нежелательное явление, возникающее при решении задач обучения по прецедентам, когда вероятность ошибки обученного алгоритма на объектах тестовой выборки оказывается существенно выше, чем средняя ошибка на обучающей выборке.

Недообучение (underfitting)— нежелательное явление, возникающее при решении задач обучения по прецедентам, когда алгоритм обучения не обеспечивает достаточно малой величины средней ошибки на обучающей выборке. Недообучение возникает при использовании недостаточно сложных моделей.

- machinelearning.ru

Наш первый метод

$$F_0 = \underset{a}{\operatorname{argmin}} \sum_i \|y_i - f(x_i, a)\|_2^2$$

$$f(x, a) = \sum_{k=0}^{p} \sum_{i=1}^{dim^k(x)} a_{ki} \prod_{u=1}^{k} x_{n(i,u)}$$

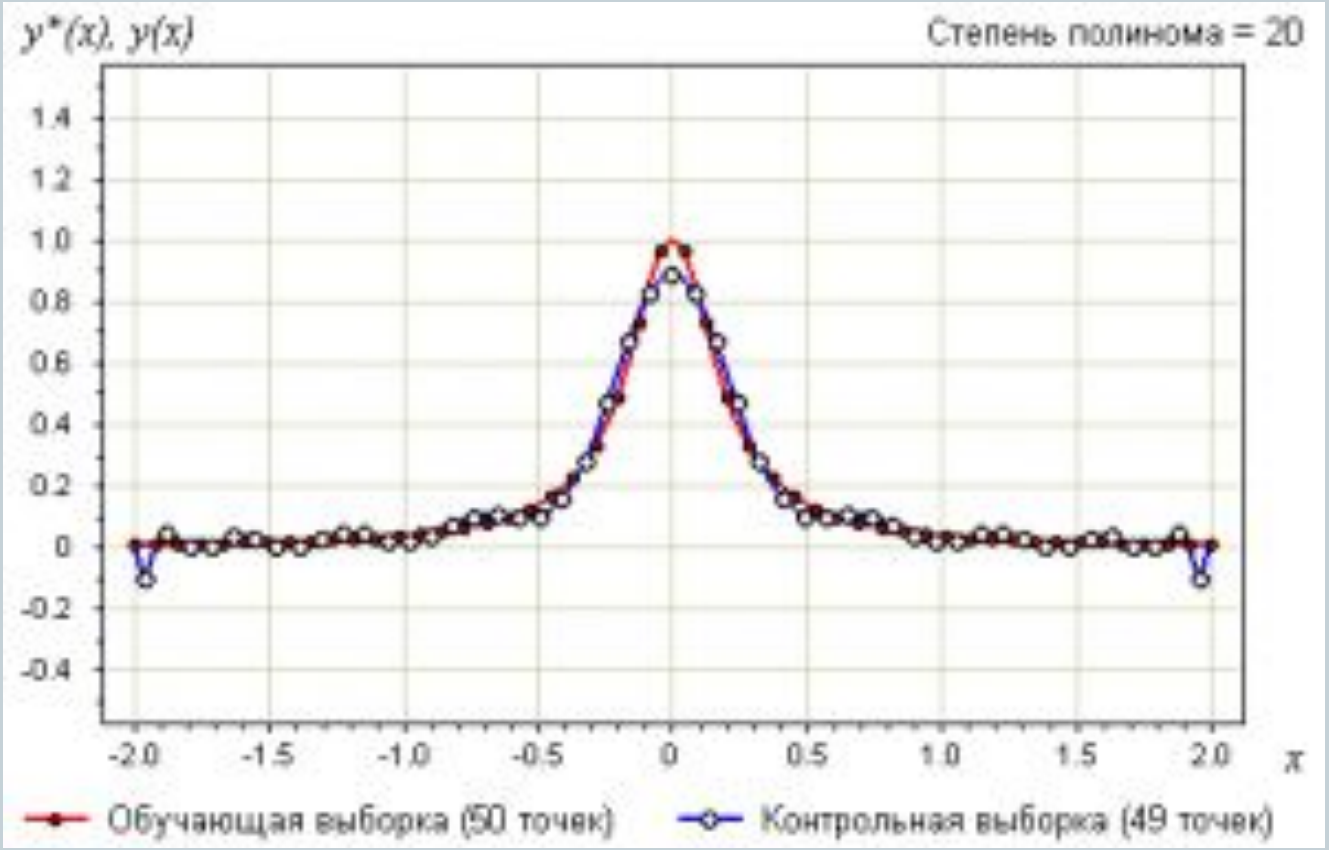
$$F_0 = \underset{a}{\operatorname{argmin}} \|y - Xa\|_2^2$$

$$a = (X^T X)^{-1} X^T y$$

Пример

По материалам machinelearning.ru

Пример



По материалам machinelearning.ru

Пример

По материалам machinelearning.ru

Задача

Дано:

$$x \sim U(0, 10]$$
$$y = ln(x)$$

Найти оптимальные p и a

Постановка в случае учителя

$$F_0 = \underset{F=A(X)}{\operatorname{argmin}}\ \mu\left(Loss(y, F(x))\right)$$

- Ожидание хотим считать по всей ген. совокупности
- Функцию обучаем на X

=> Если бы X была репрезентативной то все проще:

$$F_0 = \underset{F}{\operatorname{argmin}} \sum_{X} Loss\left(y_i, F(x_i)\right)$$

Схема тестирования

$$F_0 = \underset{F=A(X)}{\operatorname{argmin}}\ \mu\left(Loss(y, F(x))\right.$$

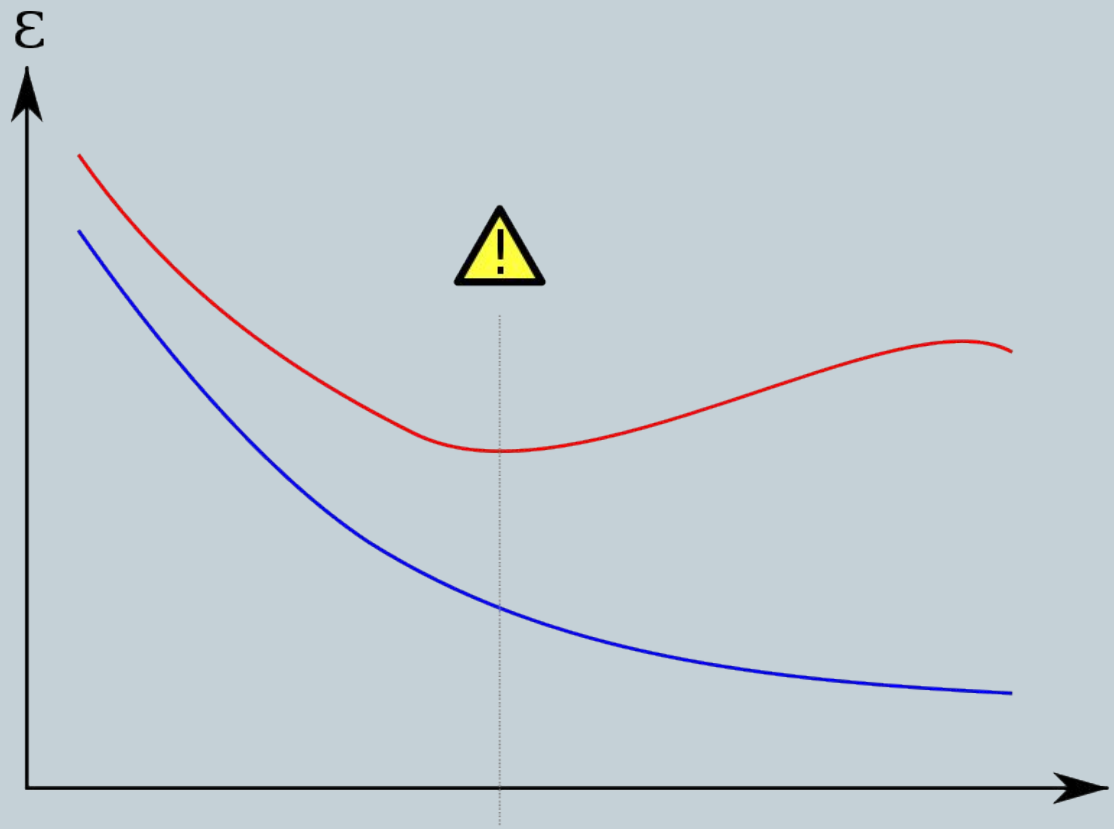
$$F_0 = \underset{F=A(X_L)}{\operatorname{argmin}}\ \mu\left(\sum_{(x,y)\in X} Loss(y, F(x))\right)$$

$$F_0 = \underset{F=A(X_L)}{\operatorname{argmin}} \sum_{(x,y)\in X_T} Loss(y, F(x))$$

$$F_0 = \underset{F=A(X_L, X_V)}{\operatorname{argmin}} \sum_{(x,y)\in X_T} Loss(y, F(x))$$

$$X = X_L \cup X_V \cup X_T$$

Overfit on validation



Как не оверфитнуться?

- White box:
  - Выбор решающего семейства при фиксированном объеме данных:
    - VC оценки
    - Оценка вероятности переобучения (по Воронцову)
    - PAC-Bayes bounds
  - Изменение процедуры подбора:
    - Игры с шагом
    - Регуляризация
- Black box:
  - Cross-validation

Теория Вапника-Червоненкиса

Владимир Наумович Вапник, Алексей Яковлевич Червоненкис

- Задача минимизации эмпирического риска

$$F_0 = \underset{F=A(X_L)}{\operatorname{argmin}} \sum_{(x,y)\in X_L} Loss(y, F(x))$$

- VC-оценка (классификация):

$$p_\epsilon \left(\sup_{F=A(X)} |p(F(x) \neq y) - p(F(x) \neq y \mid (x,y) \in X)| \geq \epsilon \right)$$

Вероятность переобучения

Воронцов Константин Вячеславович
(machinelearning.ru, ШАД в Москве)

- Вводим слабую вероятностную аксиоматику

$$\psi : X^* \to \{0, 1\}$$
$$X^L = X^l \cup \overline{X}^l$$
$$\mathbb{P}(\psi|X^L) = \frac{1}{C_L^l} \sum_{X^l} \psi(X^l)$$

- Оцениваем вероятность переобучения:

$$\delta_\mu(X^l) = p\{A(X^l)(x) \neq y | (x, y) \in \overline{X}^l\} - p\{A(X^l)(x) \neq y | (x, y) \in X^l\}$$

$$Q_\epsilon(\mu, X^L) = \mathbb{P}(\delta_\mu \geq \epsilon | X^L)$$

PAC-Bayes bounds

- Результат алгоритма – распределение над семейством
- Решающая функция –среднее выборки этого распределения

Изменение процедуры подбора

- Игры с шагом: а давайте не будем точно решать задачку

$$F_0 = \underset{F=A(X_L)}{\operatorname{argmin}} \sum_{(x,y)\in X_L} Loss(y, F(x))$$

- Поменяем Lossь так, чтобы более «рискованные» решения получали discount.

Cross-validation

- Рандомно поделим множество X на несколько кусочков
- Обучимся на одной части
- Проверим на оставшихся
- Повторим до ощущения надежности

Виды cross-validation

- 2-fold
- k-fold
- Random sub-sampling (e.g. bootstrapping)
- Leave-one-out (LOOCV)

Как принять решение по результатам CFV?

- Wilcoxon signed rank test для проверки на равенство
- Знак по выборочному среднему
- Проблемы:
  - Чем меньше выборка X тем более зависимы результаты
  - Интересно:

$$F_0 = \underset{F=A(X_L)}{\operatorname{argmin}} \ \mu \left(\sum_{(x,y)\in X} Loss(y, F(x)) \right)$$

а наблюдаем мы только 1 реализацию.

⇒ Слишком оптимистичные решения

⇒ Любое практическое исследование должно иметь эти оценки

На чем тестировать?

- Реальные данные
  - Поиск: РОМИП, TREC, Яндекс.ИМАТ, Yahoo LTRCh
  - Pascal Challenge
  - InnoCentive
- Искусственные данные (многомерный XOR)
- Задумаем «хитрое» распределение и попробуем его отгадать

Машинное Обучение: качество

ОТСЕБЯТЕНА

Решающие функции и информация

- Решающая функция несет информацию о выборке
- ⇒ Чем «короче» можно записать решающую функцию, тем меньше оверфита
- ⇒ Чем сложнее зависимость, тем больше данных надо