Онлайн обучение линейных моделей. Онлайн бутстреп.

Илья Трофимов

15.09.2016

Машинное обучение и большие данные
ФИВТ, осень 2016

Обобщенные линейные модели

$$\mathbb{E}[y|\mathbf{x}] = g^{-1}(\mathbf{w}^T\mathbf{x})$$

где $g(\cdot)$ - функция связи (link function).

Обобщенные линейные модели

$$\mathbb{E}[y|\mathbf{x}] = g^{-1}(\mathbf{w}^T\mathbf{x})$$

где $g(\cdot)$ - функция связи (link function).

- $g(z) = z$, линейная регрессия $y = \mathbf{w}^T\mathbf{x}$
- $g(z) = \ln(\frac{z}{1-z})$, логистическая регрессия
  $P(y = +1|\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}$
- $g(z) = \Phi^{-1}(z)$, пробит регрессия
  $P(y = +1|\mathbf{x}) = \Phi(\mathbf{w}^T\mathbf{x})$, где $\Phi(\cdot)$ - куммулятивная ф-ция нормального распределения
- $g(z) = \ln(z)$, Пуассоновская регрессия

$$P(y|\boldsymbol{\beta}^T\mathbf{x}) = \frac{\exp(-\lambda)\lambda^n}{n!}, \text{ где } \lambda = \exp(\boldsymbol{\beta}^T\mathbf{x})$$

Обобщенные линейные модели

Линейная модель - линейная по вектору неизвестных коэффициентов

Обобщенные линейные модели

Почему линейные модели часто используются при работе с большими данными:

- Быстрое обучение
- Быстрое предсказание
- Хорошее качество предсказания

Онлайн обучение линейных моделей

Инициализировать $\mathbf{w}$
(например $\mathbf{w} = 0$, $\mathbf{w} \sim N(0, 1)$)

Повторять:

1. Получить обучающий пример $(\mathbf{x}, y)$
2. Сделать линейное предсказание $\hat{y}_{\mathbf{w}}(x) = \mathbf{w}^T\mathbf{x}$
3. Обновить вектор $\mathbf{w}$ так, чтобы $\hat{y}_{\mathbf{w}}(x)$ стало ближе к y.

Онлайн обучение линейных моделей

Как выбирать пары $(\mathbf{x}, y)$?

- Сэмплировать из распределения $P(\mathbf{x}, y)$

Онлайн обучение линейных моделей

Как выбирать пары $(\mathbf{x}, y)$?

- Сэмплировать из распределения $P(\mathbf{x}, y)$
- Случайно из обучающей выборки, т.е. сэмплирование из эмпирического распределения, $p = 1/n$

Онлайн обучение линейных моделей

Как выбирать пары $(\mathbf{x}, y)$?

- Сэмплировать из распределения $P(\mathbf{x}, y)$
- Случайно из обучающей выборки, т.е. сэмплирование из эмпирического распределения, $p = 1/n$
- Циклически из обучающей выборки (можно читать последовательно с диска)

Онлайн обучение линейных моделей

Как выбирать пары $(\mathbf{x}, y)$?

- Сэмплировать из распределения $P(\mathbf{x}, y)$
- Случайно из обучающей выборки, т.е. сэмплирование из эмпирического распределения, $p = 1/n$
- Циклически из обучающей выборки (можно читать последовательно с диска)
  выборку нужно предварительно перемешать

Онлайн обучение линейных моделей

Как выбирать пары $(\mathbf{x}, y)$?

- Сэмплировать из распределения $P(\mathbf{x}, y)$
- Случайно из обучающей выборки, т.е. сэмплирование из эмпирического распределения, $p = 1/n$
- Циклически из обучающей выборки (можно читать последовательно с диска)
  выборку нужно предварительно перемешать
- Читать в реальном времени показания датчиков/действия пользователей/и т.п.

Обновление вектора весов

Обновление вектора весов

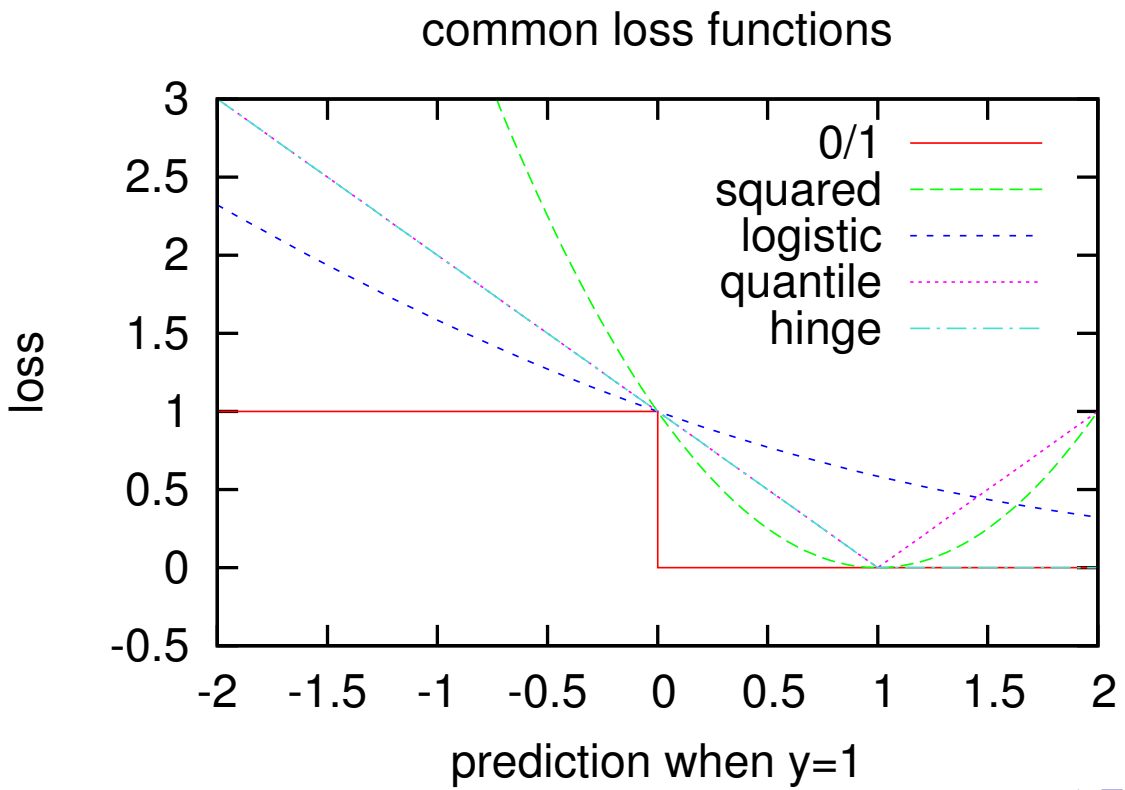
1. Выбрать подходящую функцию потерь $L(\hat{y}_{\mathbf{w}}(x), y)$.

Обновление вектора весов

1. Выбрать подходящую функцию потерь $L(\hat{y}_{\mathbf{w}}(x), y)$.
2. Обновить веса $\mathbf{w} \leftarrow \mathbf{w} - \eta \frac{\partial L(\hat{y}_{\mathbf{w}}(x), y)}{\partial \mathbf{w}}$.
   Величина η называется темпом обучения.
   Получается **метод стохастического градиента**

Обновление вектора весов

1. Выбрать подходящую функцию потерь $L(\hat{y}_{\mathbf{w}}(x), y)$.
2. Обновить веса $\mathbf{w} \leftarrow \mathbf{w} - \eta \frac{\partial L(\hat{y}_{\mathbf{w}}(x), y)}{\partial \mathbf{w}}$.
   Величина η называется темпом обучения.
   Получается **метод стохастического градиента**



Метод стохастического градиента

Инициализировать $\mathbf{w}$, $t = 1$
(например $\mathbf{w} = 0$, $\mathbf{w} \sim N(0, 1)$)

Повторять:

1. Выбрать случайный обучающий пример $(\mathbf{x}, y)$
2. Сделать линейное предсказание $\hat{y}_{\mathbf{w}}(x) = \mathbf{w}^T\mathbf{x}$
3. Обновить веса

$$\mathbf{w} \leftarrow \mathbf{w} - \eta_t \frac{\partial L(\hat{y}_{\mathbf{w}}(x), y)}{\partial \mathbf{w}}$$

4. $t = t + 1$

Сходимость

Метод стохастического градиента

$$\mathbf{w} \leftarrow \mathbf{w} - \eta_t \frac{\partial L(\hat{y}_{\mathbf{w}}(\mathbf{x}), y)}{\partial \mathbf{w}}$$

сходится к

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} L(\mathbf{w}^T \mathbf{x}_i, y_i)$$

Сходимость

Метод стохастического градиента

$$\mathbf{w} \leftarrow \mathbf{w} - \eta_t \frac{\partial L(\hat{y}_{\mathbf{w}}(\mathbf{x}), y)}{\partial \mathbf{w}}$$

сходится к

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} L(\mathbf{w}^T \mathbf{x}_i, y_i)$$

если $L(\hat{y}_{\mathbf{w}}(\mathbf{x}), y)$ - выпуклая по $\mathbf{w}$, а темп обучения убывает

$$\sum_{t=1}^{\infty} \eta_t = \infty, \qquad \sum_{t=1}^{\infty} \eta_t^2 < \infty$$

Как выбрать правильную функцию потерь?

Пусть $(\mathbf{x}, y)$ имеют совместное распределение $P(\mathbf{x}, y)$
Средний риск $Q(\mathbf{w})$ для линейной модели $\hat{y}_{\mathbf{w}}(\mathbf{x})$ равен

$$Q(\mathbf{w}) = \int_{\mathbb{R}^p} \int_{\mathbb{R}} L(\hat{y}_{\mathbf{w}}(\mathbf{x}), y) P(y, \mathbf{x}) dy \, d\mathbf{x} =$$

$$= \int_{\mathbb{R}^p} \left(\int_{\mathbb{R}} L(\hat{y}_{\mathbf{w}}(\mathbf{x}), y) P(y|\mathbf{x}) dy \right) P(\mathbf{x}) d\mathbf{x}$$

Как выбрать правильную ф-цию потерь?

Если мы хотим, чтобы минимум $Q(\mathbf{w})$ достигался в $\hat{y}_{\mathbf{w}}(\mathbf{x})$

1. $\hat{y}_{\mathbf{w}}(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$
2. $\hat{y}_{\mathbf{w}}(\mathbf{x}) =$ квантиль τ распределения $P(y|\mathbf{x})$

3. $\hat{y}_{\mathbf{w}}(\mathbf{x}) =$ медиана $P(y|\mathbf{x})$

4. Вероятность $y = 1$ (бинарная классификация)

Как выбрать правильную ф-цию потерь?

Если мы хотим, чтобы минимум $Q(\mathbf{w})$ достигался в $\hat{y}_{\mathbf{w}}(\mathbf{x})$

1. $\hat{y}_{\mathbf{w}}(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ квадратичная $L(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$
2. $\hat{y}_{\mathbf{w}}(\mathbf{x}) =$ квантиль τ распределения $P(y|\mathbf{x})$
3. $\hat{y}_{\mathbf{w}}(\mathbf{x}) =$ медиана $P(y|\mathbf{x})$
4. Вероятность $y = 1$ (бинарная классификация)

Как выбрать правильную ф-цию потерь?

Если мы хотим, чтобы минимум $Q(\mathbf{w})$ достигался в $\hat{y}_{\mathbf{w}}(\mathbf{x})$

1. $\hat{y}_{\mathbf{w}}(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ квадратичная $L(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$
2. $\hat{y}_{\mathbf{w}}(\mathbf{x}) =$ квантиль τ распределения $P(y|\mathbf{x})$
   квантильная регрессия
   $L(\hat{y}, y) = \tau(\hat{y} - y)\mathbb{I}(y \leq \hat{y}) + (1 - \tau)(y - \hat{y})\mathbb{I}(y \geq \hat{y})$
3. $\hat{y}_{\mathbf{w}}(\mathbf{x}) =$ медиана $P(y|\mathbf{x})$

4. Вероятность $y = 1$ (бинарная классификация)

Как выбрать правильную ф-цию потерь?

Если мы хотим, чтобы минимум $Q(\mathbf{w})$ достигался в $\hat{y}_{\mathbf{w}}(\mathbf{x})$

1. $\hat{y}_{\mathbf{w}}(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ квадратичная $L(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$
2. $\hat{y}_{\mathbf{w}}(\mathbf{x}) =$ квантиль τ распределения $P(y|\mathbf{x})$
   квантильная регрессия
   $L(\hat{y}, y) = \tau(\hat{y} - y)\mathbb{I}(y \leq \hat{y}) + (1 - \tau)(y - \hat{y})\mathbb{I}(y \geq \hat{y})$
3. $\hat{y}_{\mathbf{w}}(\mathbf{x}) =$ медиана $P(y|\mathbf{x})$
   квантильная регрессия с $\tau = 0.5$
   $L(\hat{y}, y) = |\hat{y} - y|$
4. Вероятность $y = 1$ (бинарная классификация)

Как выбрать правильную ф-цию потерь?

Если мы хотим, чтобы минимум $Q(\mathbf{w})$ достигался в $\hat{y}_{\mathbf{w}}(\mathbf{x})$

1. $\hat{y}_{\mathbf{w}}(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ квадратичная $L(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$
2. $\hat{y}_{\mathbf{w}}(\mathbf{x}) =$ квантиль τ распределения $P(y|\mathbf{x})$
   квантильная регрессия
   $L(\hat{y}, y) = \tau(\hat{y} - y)\mathbb{I}(y \leq \hat{y}) + (1 - \tau)(y - \hat{y})\mathbb{I}(y \geq \hat{y})$
3. $\hat{y}_{\mathbf{w}}(\mathbf{x}) =$ медиана $P(y|\mathbf{x})$
   квантильная регрессия с $\tau = 0.5$
   $L(\hat{y}, y) = |\hat{y} - y|$
4. Вероятность $y = 1$ (бинарная классификация)
   логистическая $L(\hat{y}, y) = \log(1 + \exp(-y\hat{y}))$

Как выбрать правильную ф-цию потерь?

Если мы хотим, чтобы минимум $Q(\mathbf{w})$ достигался в $\hat{y}_{\mathbf{w}}(\mathbf{x})$

1. $\hat{y}_{\mathbf{w}}(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ квадратичная $L(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$
2. $\hat{y}_{\mathbf{w}}(\mathbf{x}) =$ квантиль τ распределения $P(y|\mathbf{x})$
   квантильная регрессия
   $L(\hat{y}, y) = \tau(\hat{y} - y)\mathbb{I}(y \leq \hat{y}) + (1 - \tau)(y - \hat{y})\mathbb{I}(y \geq \hat{y})$
3. $\hat{y}_{\mathbf{w}}(\mathbf{x}) =$ медиана $P(y|\mathbf{x})$
   квантильная регрессия с $\tau = 0.5$
   $L(\hat{y}, y) = |\hat{y} - y|$
4. Вероятность $y = 1$ (бинарная классификация)
   логистическая $L(\hat{y}, y) = \log(1 + \exp(-y\hat{y}))$
5. ...или $L(\hat{y}, y)$ - минус лог-правдоподобие обобщенной линейной модели

Как выбрать правильную ф-цию потерь?

1. Типичная цена квартиры в районе?

2. Прибыль от продажи акций?
3. Вероятность клика по онлайн рекламе?

Как выбрать правильную ф-цию потерь?

1. Типичная цена квартиры в районе?
   квантильная регрессия $\tau = 0.5$
2. Прибыль от продажи акций?
3. Вероятность клика по онлайн рекламе?

Как выбрать правильную ф-цию потерь?

1. Типичная цена квартиры в районе?
   квантильная регрессия $\tau = 0.5$
2. Прибыль от продажи акций? квадратичная
3. Вероятность клика по онлайн рекламе?

Как выбрать правильную ф-цию потерь?

1. Типичная цена квартиры в районе?
   квантильная регрессия $\tau = 0.5$
2. Прибыль от продажи акций? квадратичная
3. Вероятность клика по онлайн рекламе? логистическая

Как измерить качество предсказания?

Инициализировать $\mathbf{w}, t = 1$

Повторять:

1. Получить обучающий пример $(\mathbf{x}, y)$
2. Сделать линейное предсказание $\hat{y}_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$.
3. Измерить точность прогноза $E_t = L(\hat{y}_{\mathbf{w}}(\mathbf{x}), y)$
4. Обновить вектор весов так, чтобы $\hat{y}_{\mathbf{w}}(\mathbf{x})$ стало ближе к y.
5. $t = t + 1$

Как измерить качество предсказания?

Инициализировать $\mathbf{w}, t = 1$
Повторять:

1. Получить обучающий пример $(\mathbf{x}, y)$
2. Сделать линейное предсказание $\hat{y}_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$.
3. Измерить точность прогноза $E_t = L(\hat{y}_{\mathbf{w}}(\mathbf{x}), y)$
4. Обновить вектор весов так, чтобы $\hat{y}_{\mathbf{w}}(\mathbf{x})$ стало ближе к y.
5. $t = t + 1$

Техника "**Progressive Validation (PV)**"

Ошибка PV равна $Err = \frac{1}{T}\sum_{t=1}^{T} E_t$

Как измерить качество предсказания?

Инициализировать $\mathbf{w}, t = 1$
Повторять:

1. Получить обучающий пример $(\mathbf{x}, y)$
2. Сделать линейное предсказание $\hat{y}_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$.
3. Измерить точность прогноза $E_t = L(\hat{y}_{\mathbf{w}}(\mathbf{x}), y)$
4. Обновить вектор весов так, чтобы $\hat{y}_{\mathbf{w}}(\mathbf{x})$ стало ближе к y.
5. $t = t + 1$

Техника "**Progressive Validation (PV)**"

Ошибка PV равна $Err = \frac{1}{T}\sum_{t=1}^{T} E_t$

Корректно считается только при однократном проходе по выборке!

Трюки для онлайн обучения

1. Учет весов обучающих примеров
2. Адаптивный темп обучения
3. Адаптивная нормализация признаков

Обучение с весам примеров

Типичная ситуация: необходимо обучить классификатор, но ошибки 1-го и 2-го рода не равноценны.

Пример: в обнаружении спама, предсказать, что хорошее письмо - это спам хуже, чем спам - это хорошее письмо.

Обучение с весам примеров

Типичная ситуация: необходимо обучить классификатор, но ошибки 1-го и 2-го рода не равноценны.

Пример: в обнаружении спама, предсказать, что хорошее письмо - это спам хуже, чем спам - это хорошее письмо.

Пусть обучающий пример имеет вес l (Importance).
Как нужно модифицировать правильно обновления вектора w (метод стохастического градиента) с учетом l?

Обучение с весам примеров

Типичная ситуация: необходимо обучить классификатор, но ошибки 1-го и 2-го рода не равноценны.

Пример: в обнаружении спама, предсказать, что хорошее письмо - это спам хуже, чем спам - это хорошее письмо.

Пусть обучающий пример имеет вес I (Importance).
Как нужно модифицировать правильно обновления вектора w (метод стохастического градиента) с учетом I?

"Наивный подход": $\mathbf{w} \leftarrow \mathbf{w} - \eta I \frac{\partial L(\hat{y}_{\mathbf{w}}(\mathbf{x}), y)}{\partial \mathbf{w}}$.

Обучение с весам примеров

Задача минимизации эмпирического риска

$$\mathbf{w} = \operatorname*{argmin}_{\mathbf{w}} Q(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^{n} l_i L(\hat{y}_{\mathbf{w}}(\mathbf{x}_i), y_i)$$

Добавление весов эквивалентно копированию примера i-го примера l_i раз.

Обучение с весам примеров

Задача минимизации эмпирического риска

$$\mathbf{w} = \operatorname*{argmin}_{\mathbf{w}} Q(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^{n} l_i L(\hat{y}_{\mathbf{w}}(\mathbf{x}_i), y_i)$$

Добавление весов эквивалентно копированию примера i-го примера l_i раз.

Идея: подобрать функцию $s(\eta, \mathbf{x}, y)$ такую, что обновление, сделанное l раз

$$\mathbf{w} \leftarrow \mathbf{w} - s(\eta, \mathbf{x}, y)$$

было бы эквивалентно

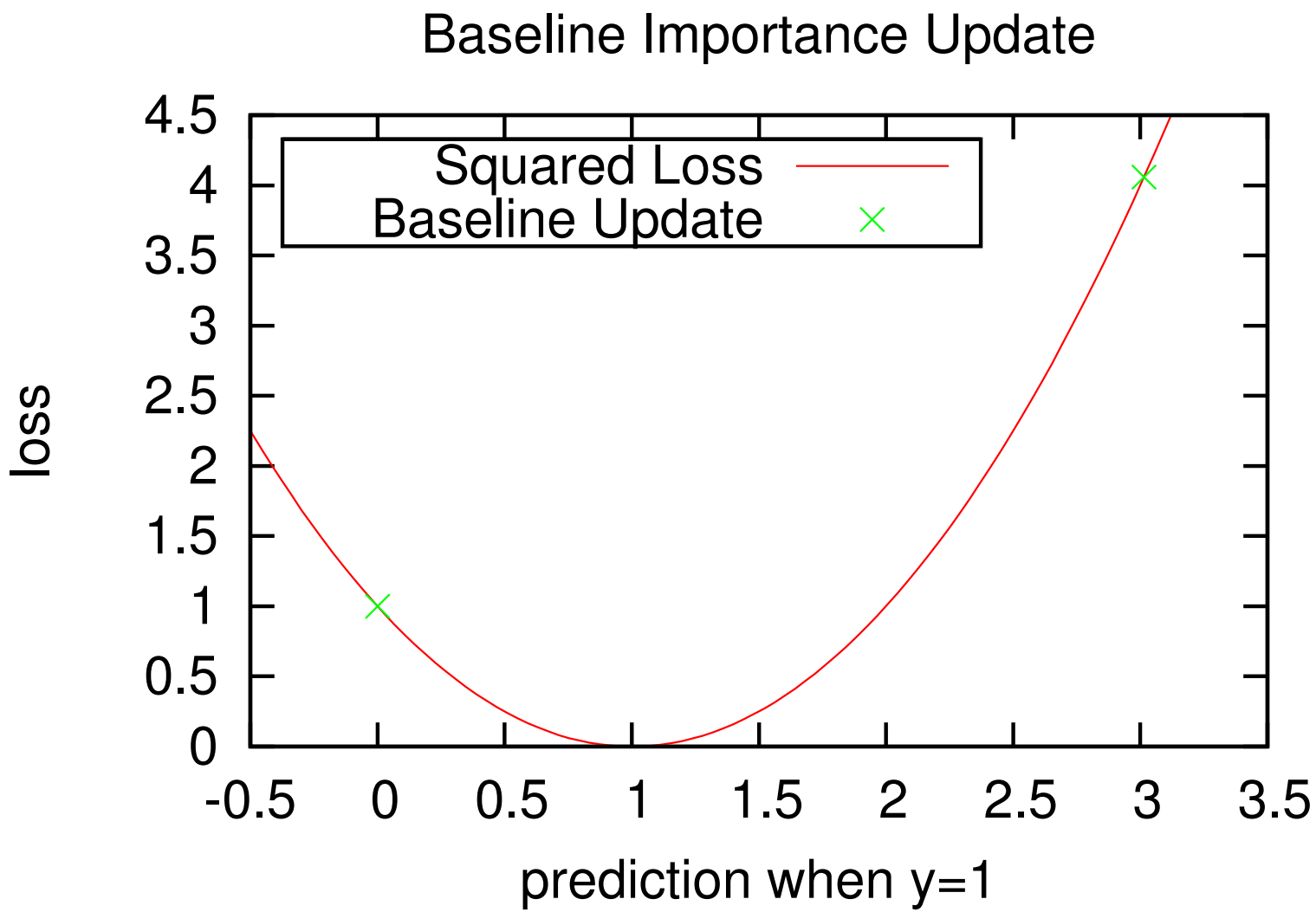
$$\mathbf{w} \leftarrow \mathbf{w} - s(\eta l, \mathbf{x}, y)$$

Учет важности примеров

Loss	$\ell(p, y)$	Invariant Update $s(h)$
Squared	$\frac{1}{2}(y-p)^2$	$\frac{p-y}{x^\top x}\left(1-e^{-h\eta x^\top x}\right)$
Logistic	$\log(1+e^{-yp})$	$\frac{W(e^{h\eta x^\top x+yp+e^{yp}})-h\eta x^\top x-e^{yp}}{yx^\top x}$ for $y\in\{-1,1\}$
Exponential	e^{-yp}	$\frac{py-\log(h\eta x^\top x+e^{py})}{x^\top xy}$ for $y\in\{-1,1\}$
Logarithmic	$y\log\frac{y}{p}+(1-y)\log\frac{1-y}{1-p}$	if $y=0$ $\frac{p-1+\sqrt{(p-1)^2+2h\eta x^\top x}}{x^\top x}$ if $y=1$ $\frac{p-\sqrt{p^2+2h\eta x^\top x}}{x^\top x}$
Hellinger	$2(1-\sqrt{py}-\sqrt{(1-p)(1-y)})$	if $y=0$ $\frac{p-1+\frac{1}{4}(12h\eta x^\top x+8(1-p)^{3/2})^{2/3}}{x^\top x}$ if $y=1$ $\frac{p-\frac{1}{4}(12h\eta x^\top x+8p^{3/2})^{2/3}}{x^\top x}$
Hinge	$\max(0,1-yp)$	$-y\min\left(h\eta,\frac{1-yp}{x^\top x}\right)$ for $y\in\{-1,1\}$
τ-Quantile	if $y>p$ $\tau(y-p)$ if $y\le p$ $(1-\tau)(p-y)$	if $y>p$ $-\tau\min(h\eta,\frac{y-p}{\tau x^\top x})$ if $y\le p$ $(1-\tau)\min(h\eta,\frac{p-y}{(1-\tau)x^\top x})$

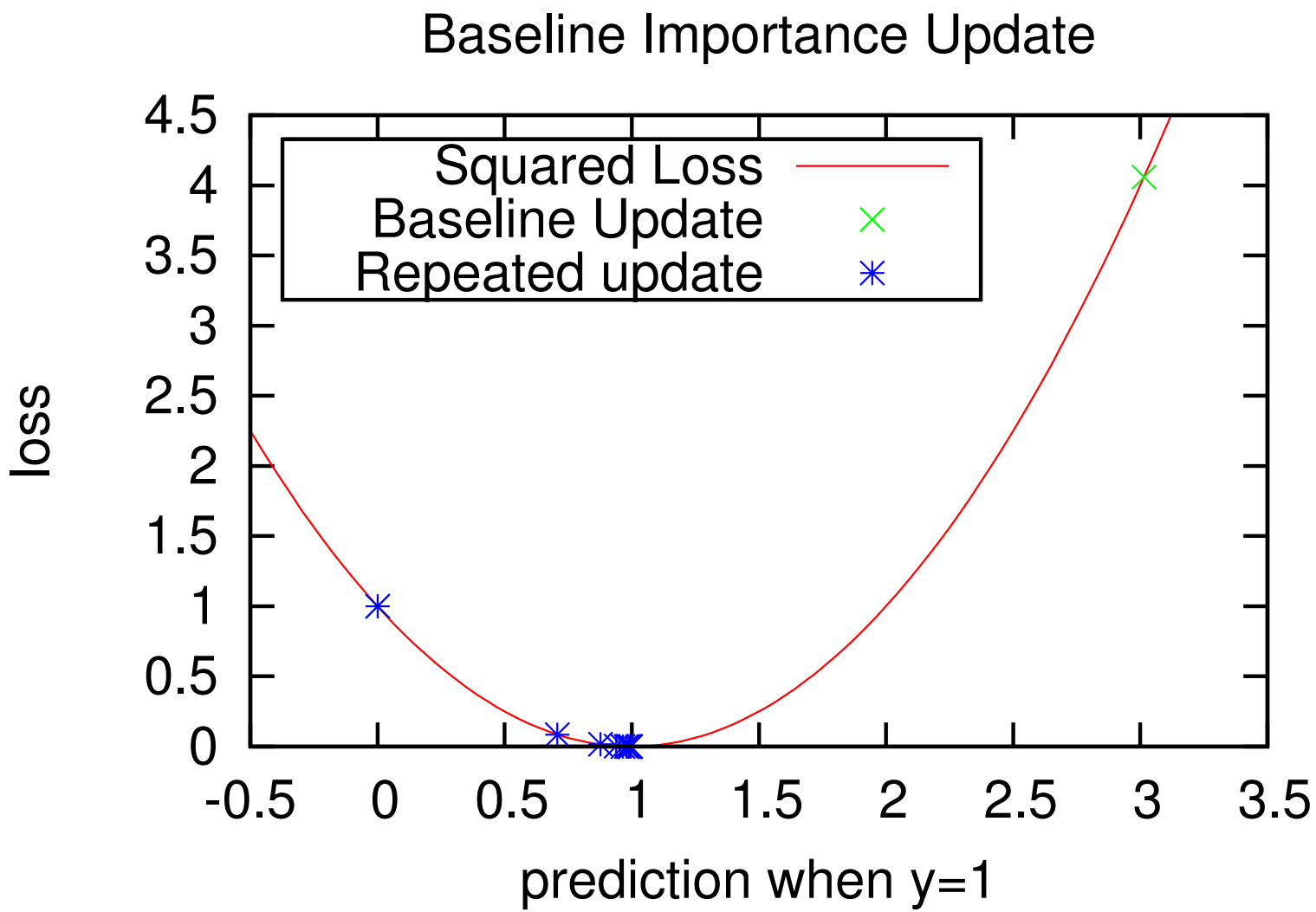
Учет важности примеров

$\mathbf{w} \leftarrow \mathbf{w} - \eta I \frac{\partial L(\hat{y}_{\mathbf{w}}(\mathbf{x}), y)}{\partial \mathbf{w}}$ работает плохо.



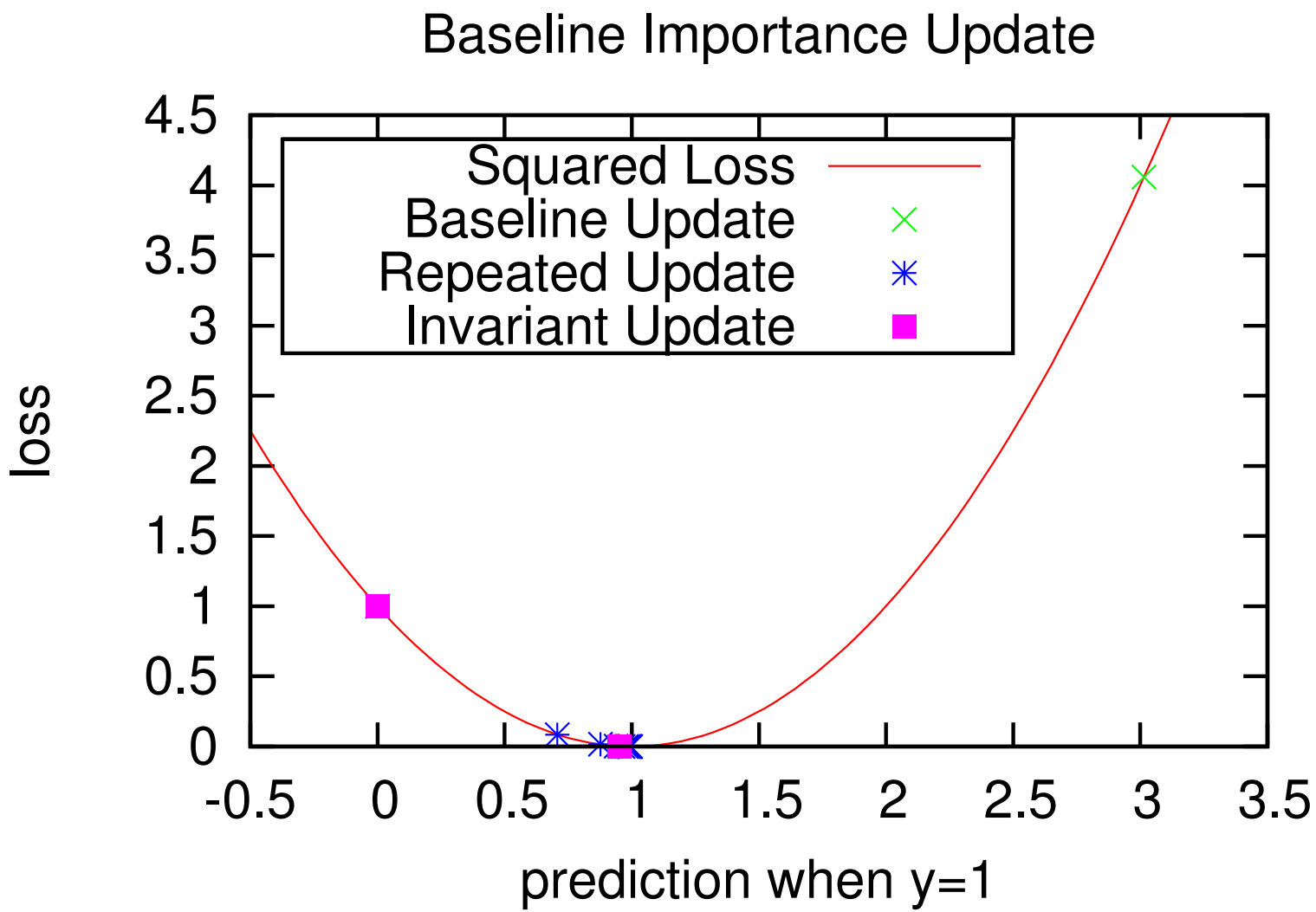
Учет важности примеров

работает лучше: $\mathbf{w} \leftarrow \mathbf{w} - \eta \frac{\partial L(\hat{y}_{\mathbf{w}}(\mathbf{x}), y)}{\partial \mathbf{w}}$ повтор I раз



Учет важности примеров

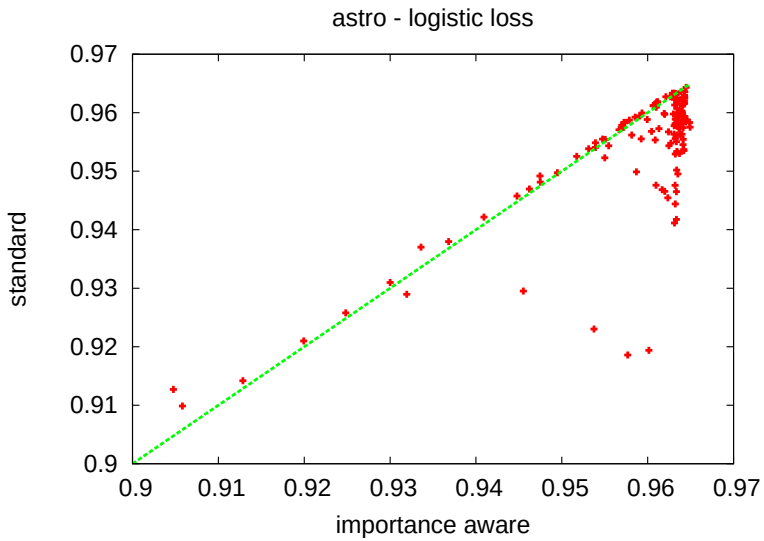
работает еще лучше: $\mathbf{w} \leftarrow \mathbf{w} - s(\eta I)\frac{\partial L(\hat{y}_{\mathbf{w}}(\mathbf{x}),y)}{\partial \mathbf{w}}$

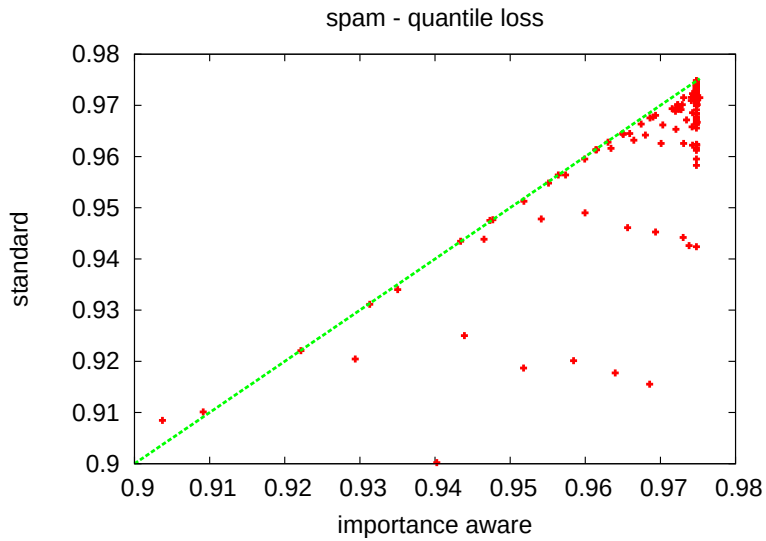


Устойчивость для задач с одинаковой важностью примеров

Для датасетов astro, spam произведен перебор темпов обучения, обучение и тестирование сделано для случая обновлений по

- правилу стохастического градиента
- инвариантные обновления с ф-цией $s(\eta, \mathbf{x}, y)$





Адаптивный темп обучения

Темп обучения должен падать, но с какой скоростью?

Адаптивный темп обучения

Темп обучения должен падать, но с какой скоростью?

- $\eta_t = 1/t^\alpha$, где $\alpha \in [0.5, 1.0]$, если обучаться на конечной выборке
- $\eta_t = \eta_0$, если обучаться на потоковых данных

Адаптивный темп обучения

Темп обучения должен падать, но с какой скоростью?

- $\eta_t = 1/t^\alpha$, где $\alpha \in [0.5, 1.0]$, если обучаться на конечной выборке
- $\eta_t = \eta_0$, если обучаться на потоковых данных

AdaGrad

Обозначим $g_{it} = \frac{\partial L(\hat{y}_{\mathbf{w}}(\mathbf{x}_t), y_t)}{\partial w_i}$

Адаптивный темп обучения

Темп обучения должен падать, но с какой скоростью?

- $\eta_t = 1/t^\alpha$, где $\alpha \in [0.5, 1.0]$, если обучаться на конечной выборке
- $\eta_t = \eta_0$, если обучаться на потоковых данных

AdaGrad

Обозначим $g_{it} = \frac{\partial L(\hat{y}_{\mathbf{w}}(\mathbf{x}_t), y_t)}{\partial w_i}$

Адаптивный темп обучения: $w_i \leftarrow w_i - \eta_i g_{it}$

Адаптивный темп обучения

Темп обучения должен падать, но с какой скоростью?

- $\eta_t = 1/t^\alpha$, где $\alpha \in [0.5, 1.0]$, если обучаться на конечной выборке
- $\eta_t = \eta_0$, если обучаться на потоковых данных

AdaGrad

Обозначим $g_{it} = \frac{\partial L(\hat{y}_{\mathbf{w}}(\mathbf{x}_t), y_t)}{\partial w_i}$

Адаптивный темп обучения: $w_i \leftarrow w_i - \eta_i g_{it}$

где $\eta_i = \frac{\eta}{\sqrt{\sum_{t'=1}^{t} g_{it'}^2}}$ темп обучения для признака i

Адаптивный темп обучения

Адаптивный темп обучения: $w_i \leftarrow w_i - \eta \frac{g_{it}}{\sqrt{\sum_{t'=1}^{t} g_{it'}^2}}$

Адаптивный темп обучения

Адаптивный темп обучения: $w_i \leftarrow w_i - \eta \frac{g_{it}}{\sqrt{\sum_{t'=1}^{t} g_{it'}^2}}$

$$g_{it} = \frac{\partial L(\hat{y}_{\mathbf{w}}(\mathbf{x}_t), y_t)}{\partial w_i} = \left.\frac{\partial L(p, y_t)}{\partial p}\right|_{p=\hat{y}_{\mathbf{w}}(\mathbf{x}_t)} \frac{\partial \hat{y}_{\mathbf{w}}(\mathbf{x}_t)}{\partial w_i}$$

$$= \left.\frac{\partial L(p, y_t)}{\partial p}\right|_{p=\hat{y}_{\mathbf{w}}(\mathbf{x}_t)} \left.\frac{\partial g^{-1}(s)}{\partial s}\right|_{s=\mathbf{w}^T\mathbf{x}_t} x_{it}$$

Адаптивный темп обучения

Адаптивный темп обучения: $w_i \leftarrow w_i - \eta \frac{g_{it}}{\sqrt{\sum_{t'=1}^{t} g_{it'}^2}}$

$$g_{it} = \frac{\partial L(\hat{y}_{\mathbf{w}}(\mathbf{x}_t), y_t)}{\partial w_i} = \left.\frac{\partial L(p, y_t)}{\partial p}\right|_{p=\hat{y}_{\mathbf{w}}(\mathbf{x}_t)} \frac{\partial \hat{y}_{\mathbf{w}}(\mathbf{x}_t)}{\partial w_i}$$

$$= \left.\frac{\partial L(p, y_t)}{\partial p}\right|_{p=\hat{y}_{\mathbf{w}}(\mathbf{x}_t)} \left.\frac{\partial g^{-1}(s)}{\partial s}\right|_{s=\mathbf{w}^T\mathbf{x}_t} x_{it}$$

Адаптивный темп обучения

Адаптивный темп обучения: $w_i \leftarrow w_i - \eta \frac{g_{it}}{\sqrt{\sum_{t'=1}^{t} g_{it'}^2}}$

$$g_{it} = \frac{\partial L(\hat{y}_{\mathbf{w}}(\mathbf{x}_t), y_t)}{\partial w_i} = \left.\frac{\partial L(p, y_t)}{\partial p}\right|_{p=\hat{y}_{\mathbf{w}}(\mathbf{x}_t)} \frac{\partial \hat{y}_{\mathbf{w}}(\mathbf{x}_t)}{\partial w_i}$$

$$= \left.\frac{\partial L(p, y_t)}{\partial p}\right|_{p=\hat{y}_{\mathbf{w}}(\mathbf{x}_t)} \left.\frac{\partial g^{-1}(s)}{\partial s}\right|_{s=\mathbf{w}^T\mathbf{x}_t} x_{it}$$

Значения часто встречающихся признаков быстро стабилизируются, по ним нужно делать маленькие шаги. Для редко встречающихся признаков эффективнее делать большие шаги.

Визуализация онлайн обучения

http://sebastianruder.com/content/images/2016/01/contours_evaluation_optimizers.gif

http://sebastianruder.com/content/images/2016/01/saddle_point_evaluation_optimizers.gif

Учет размерности переменных

Шаг стохастического градиента: $w_i \leftarrow w_i - \eta g_{it}$

$$g_{it} = \left.\frac{\partial L(p, y_t)}{\partial p}\right|_{p=\hat{y}_{\mathbf{w}}(\mathbf{x}_t)} \left.\frac{\partial g^{-1}(s)}{\partial s}\right|_{s=\mathbf{w}^T\mathbf{x}_t} x_{it}$$

$$w_i \leftarrow w_i - C x_{it}$$

Учет размерности переменных

Шаг стохастического градиента: $w_i \leftarrow w_i - \eta g_{it}$

$$g_{it} = \left.\frac{\partial L(p, y_t)}{\partial p}\right|_{p=\hat{y}_{\mathbf{w}}(\mathbf{x}_t)} \left.\frac{\partial g^{-1}(s)}{\partial s}\right|_{s=\mathbf{w}^T\mathbf{x}_t} x_{it}$$

$$w_i \leftarrow w_i - C x_{it}$$

Проблема: умножение x_i на 2 должно влечь за собой уменьшение w_i в 2 раза, чтобы предсказание не менялось.
$\Rightarrow$ В датасетах признаки имеют разные масштабы!

Нормализация признаков?

Для каждого признака x_i вычислить:
Выборочное среднее $\mu_i = \sum_{t=1}^{n} x_{it}$
Выборочное стд. отклонение $\sigma_i = \sqrt{\sum_{t=1}^{n} (x_{it} - \mu_i)^2}$
$x_i' \leftarrow \frac{x_i - \mu_i}{\sigma_i}$.

Нормализация признаков?

Для каждого признака x_i вычислить:
Выборочное среднее $\mu_i = \sum_{t=1}^{n} x_{it}$
Выборочное стд. отклонение $\sigma_i = \sqrt{\sum_{t=1}^{n} (x_{it} - \mu_i)^2}$
$x_i' \leftarrow \frac{x_i - \mu_i}{\sigma_i}$.
Проблемы:

1. Не работает онлайн.
2. Выборка перестает быть разреженной.

Адаптивный учет масштаба переменных

Алгоритм NG(темп обучения η)

1. Инициализировать $w_i = 0$, t $= 0$, $s_i = 0$, $N = 0$
2. Получить обучающий пример $(\mathbf{x}, y)$
   1. Для каждого i, если $|x_i| > s_i$
      1. $w_i \leftarrow \frac{w_i s_i^2}{|x_i|^2}$
      2. $s_i \leftarrow |x_i|$
   2. $\hat{y}_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$
   3. $N \leftarrow N + \sum_i \frac{x_i^2}{s_i^2}$
   4. Для каждого i,
      1. $w_i \leftarrow w_i - \eta \sqrt{\frac{t}{N}} \frac{1}{s_i^2} \frac{\partial L(\hat{y}, y)}{\partial w_i}$
3. t $=$ t $+ 1$

Адаптивный учет масштаба переменных

Алгоритм NG(темп обучения η)

1. Инициализировать $w_i = 0$, t $= 0$, $s_i = 0$, $N = 0$
2. Получить обучающий пример $(\mathbf{x}, y)$
   1. Для каждого i, если $|x_i| > s_i$
      1. Перевести w_i в новый масштаб
      2. Изменить масштаб
   2. $\hat{y}_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
   3. Изменить общий масштаб
   4. Для каждого i,
      1. $w_i \leftarrow w_i - \eta \sqrt{\frac{t}{N}} \frac{1}{s_i^2} \frac{\partial L(\hat{y}, y)}{\partial w_i}$
3. t $=$ t $+ 1$

Онлайн бутстреп

Обучающая выборка:

$$\mathbb{Z} = (z_1, z_2, \ldots, z_n), \quad z_i = (\mathbf{x}_i, y_i)$$

Построим B выборок из $\mathbb{Z}$, используя случайный выбор с возвращением.

Онлайн бутстреп

Обучающая выборка:

$$\mathbb{Z} = (z_1, z_2, \ldots, z_n), \quad z_i = (\mathbf{x}_i, y_i)$$

Построим B выборок из $\mathbb{Z}$, используя случайный выбор с возвращением.

Исходная выборка

1	2	3	4	5

Онлайн бутстреп

Обучающая выборка:

$$\mathbb{Z} = (z_1, z_2, \ldots, z_n), \quad z_i = (\mathbf{x}_i, y_i)$$

Построим B выборок из $\mathbb{Z}$, используя случайный выбор с возвращением.

Исходная выборка

1	2	3	4	5

Выборки после бутстрепа, B = 4

1	1	2	3	5
1	2	2	3	5
2	3	3	4	4
1	2	3	4	5

Онлайн бутстреп

Обучающая выборка:

$$\mathbb{Z} = (z_1, z_2, \ldots, z_n), \quad z_i = (\mathbf{x}_i, y_i)$$

Построим B выборок из $\mathbb{Z}$, используя случайный выбор с возвращением.

Исходная выборка

1	2	3	4	5

Выборки после бутстрепа, B = 4

1	1	2	3	5
1	2	2	3	5
2	3	3	4	4
1	2	3	4	5

P.S. рекомендуется брать $B = 10000$

Бутстреп

Идея бутстрепа: $(\mathbf{x}, y) \sim P(\mathbf{x}, y)$, но $P(\mathbf{x}, y)$ - неизвестно!
Тогда вместо $P(\mathbf{x}, y)$ будем использовать *эмпирическое* распределение, возвращающее $(\mathbf{x}_i, y_i)$ с вероятностью $\frac{1}{n}$.

Бутстреп

Идея бутстрепа: $(\mathbf{x}, y) \sim P(\mathbf{x}, y)$, но $P(\mathbf{x}, y)$ - неизвестно!
Тогда вместо $P(\mathbf{x}, y)$ будем использовать *эмпирическое* распределение, возвращающее $(\mathbf{x}_i, y_i)$ с вероятностью $\frac{1}{n}$.

Пусть $S(\mathbb{Z})$ - величина, рассчитанная на основе выборки $\mathbb{Z}$.

Бутстреп

Идея бутстрепа: $(\mathbf{x}, y) \sim P(\mathbf{x}, y)$, но $P(\mathbf{x}, y)$ - неизвестно!
Тогда вместо $P(\mathbf{x}, y)$ будем использовать *эмпирическое* распределение, возвращающее $(\mathbf{x}_i, y_i)$ с вероятностью $\frac{1}{n}$.

Пусть $S(\mathbb{Z})$ - величина, рассчитанная на основе выборки $\mathbb{Z}$.
Примеры:

- параметры регрессии
- предсказание в фиксированной точке

Бутстреп

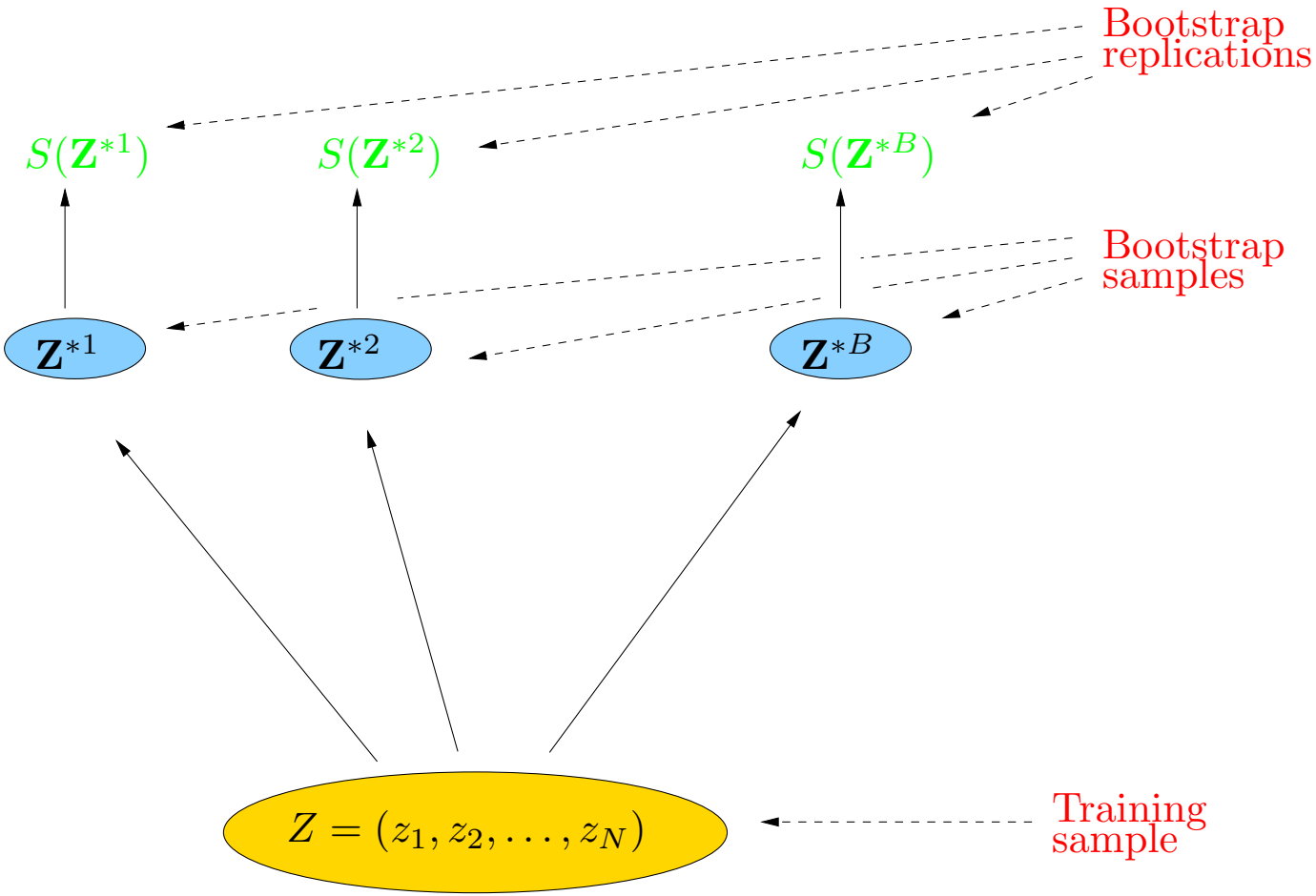
Идея бутстрепа: $(\mathbf{x}, y) \sim P(\mathbf{x}, y)$, но $P(\mathbf{x}, y)$ - неизвестно! Тогда вместо $P(\mathbf{x}, y)$ будем использовать *эмпирическое* распределение, возвращающее $(\mathbf{x}_i, y_i)$ с вероятностью $\frac{1}{n}$.

Пусть $S(\mathbb{Z})$ - величина, рассчитанная на основе выборки $\mathbb{Z}$. Примеры:

- параметры регрессии
- предсказание в фиксированной точке

Bagging (bootstrap aggregating): для предсказания можно использовать среднее или голосование моделей, обученных на $\mathbb{Z}^{*b}, b = 1 \dots B$.

Бутстреп



Hastie, T., Tibshirani, R., Friedman, J. (2005). The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 27(2), 83-85.

Бутстреп

$$\bar{S}(\mathbb{Z}^*) = \frac{1}{B}\sum_{b=1}^{B} S(\mathbb{Z}^{*b})$$

$$Var[S(\mathbb{Z})] = \frac{1}{B-1}\sum_{b=1}^{B}(S(\mathbb{Z}^{*b}) - \bar{S}(\mathbb{Z}^*))^2$$

Бутстреп

Доверительный интервал.
Отсортируем $\mathbb{Z}^{*b}$, где $b = 1 \dots B$ по возрастанию:

$$S(\mathbb{Z}^{*b_1}) \leq S(\mathbb{Z}^{*b_2}) \leq \dots \leq S(\mathbb{Z}^{*b_B})$$

Если нужен $1 - \alpha$ доверительный интервал, то нужно откинуть $\frac{\alpha}{2}$ точек в начале и конце списка, получим искомый интервал:

$$\left(S(\mathbb{Z}^{*b_c}), S(\mathbb{Z}^{*b_d})\right)$$

$$c = \left[\frac{\alpha}{2}B\right]$$

$$d = \left[(1 - \frac{\alpha}{2})B\right]$$

Онлайн бутстреп

Бутстреп неудобно делать на больших выборках.

Онлайн бутстреп

Бутстреп неудобно делать на больших выборках. Онлайн?

Онлайн бутстреп

Бутстреп неудобно делать на больших выборках. Онлайн?

Какова вероятность, что пример z_i будет в бутстрапированной выборке k раз?

Онлайн бутстреп

Бутстреп неудобно делать на больших выборках. Онлайн?

Какова вероятность, что пример z_i будет в бутстрапированной выборке k раз?

$$Binom(k|n,p) = C_n^k p^k (1-p)^{n-k}, \text{ где } p = 1/n$$

Онлайн бутстреп

Бутстреп неудобно делать на больших выборках. Онлайн?

Какова вероятность, что пример z_i будет в бутстрапированной выборке k раз?

$$Binom(k|n,p) = C_n^k p^k (1-p)^{n-k}, \text{ где } p = 1/n$$

При $n \to +\inf, pn = const$ биномиальное распределение переходит в распределение Пуассона

$$Binom(k|n,p) \to Poisson(k|pn) = \frac{p^k \exp(-p)}{k!}$$

Онлайн бутстреп

Бутстреп неудобно делать на больших выборках. Онлайн?

Какова вероятность, что пример z_i будет в бутстрапированной выборке k раз?

$$Binom(k|n,p) = C_n^k p^k (1-p)^{n-k}, \text{ где } p = 1/n$$

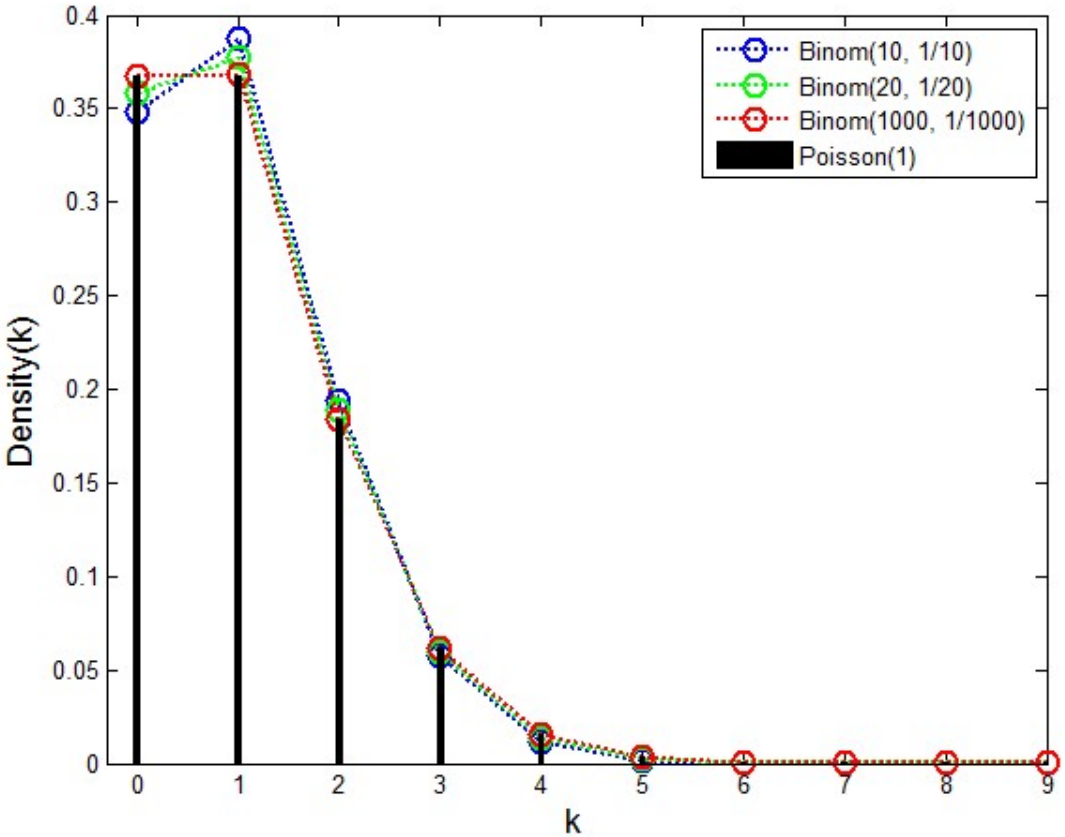
При $n \to +\inf, pn = const$ биномиальное распределение переходит в распределение Пуассона

$$Binom(k|n,p) \to Poisson(k|pn) = \frac{p^k \exp(-p)}{k!}$$

Так как $pn = 1$

$$Binom(k|n,1/n) \to Poisson(k|1) = \frac{\exp(-1)}{k!}$$

Онлайн бутстреп



Qin, Z., Petricek, V., & Karampatziakis, N. (2013). Efficient Online Bootstrapping for Large Scale Learning. arXiv Preprint `http://arxiv.org/abs/1312.5021`

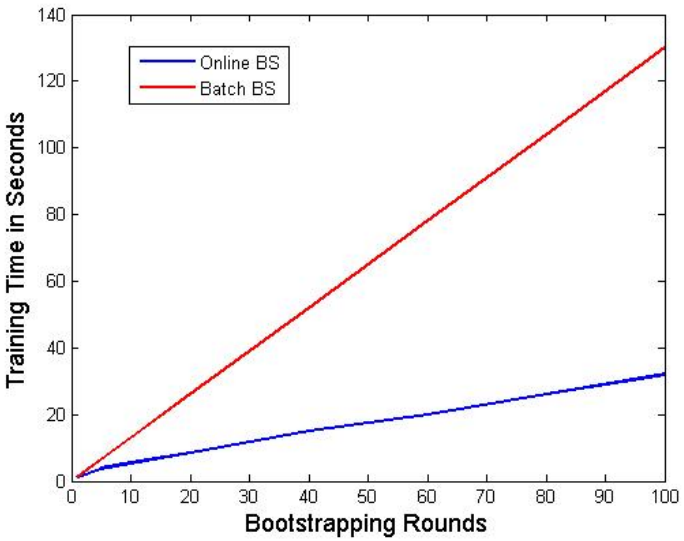
Онлайн бутстреп

Повторять:

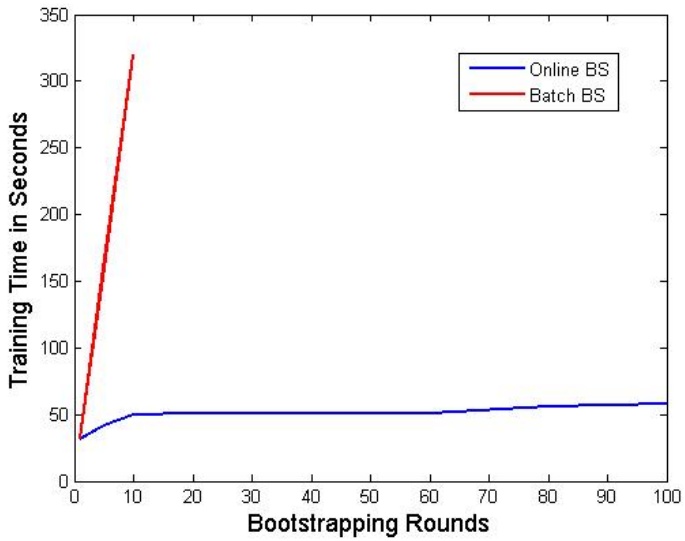
1. Получить обучающий пример $(\mathbf{x}, y)$
2. Получить вес примера I.
3. Цикл $b = 1 \dots B$:
   1. Сделать линейное предсказание $\hat{y}_{w^b}(x) = \sum_i w_i^b x_i$.
   2. $k \sim Poisson(1)$
   3. $I \leftarrow Ik$
   4. обновить вектор весов w^b чтобы $\hat{y}_{w^b}(x)$ стало ближе к y.

Вернуть : $w^b,\ b = 1 \dots B$

Онлайн бутстреп



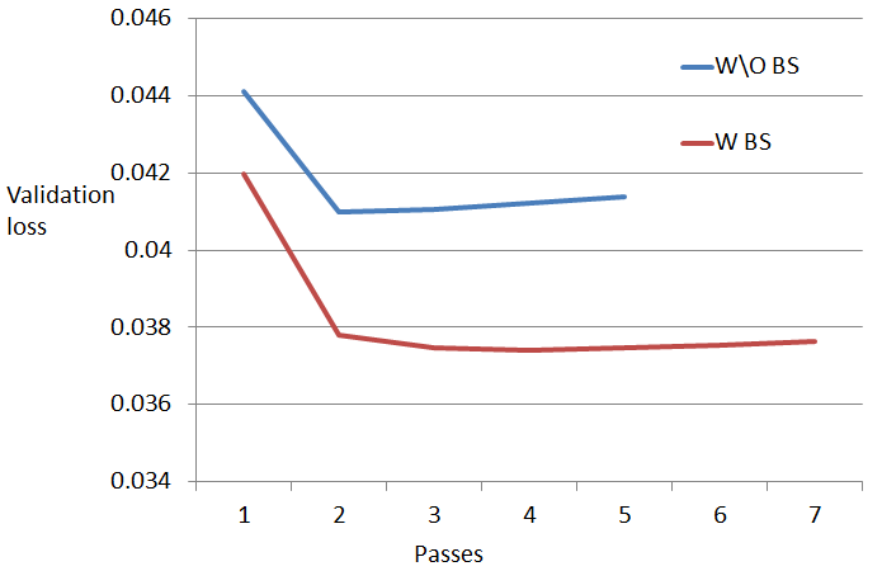
75K Dataset



RCV1 Training

Qin, Z., Petricek, V., & Karampatziakis, N. (2013). Efficient Online Bootstrapping for Large Scale Learning. arXiv Preprint `http://arxiv.org/abs/1312.5021`

Онлайн бутстреп



Method	Error Rate
Base Learner (BL)	6.01%
BL + Online BS (N=20)	5.37%
Tuned Learner (TL)	4.64%
TL + Online BS (N=4)	4.58%

Улучшение качества от техники bagging для классификации датасета RCV1. Progressive validation error и ошибка на тесте.

Qin, Z., Petricek, V., & Karampatziakis, N. (2013). Efficient Online Bootstrapping for Large Scale Learning. arXiv Preprint `http://arxiv.org/abs/1312.5021`

Список литературы

[VW] Vowpal Wabbit project, `http://hunch.net/~vw`, 2007-2012.

[Quantile Regression] Roger Koenker, Quantile Regression, Econometric Society Monograph Series, Cambridge University Press, 2005.

[Classification Consistency] Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. COLT, 2005.

[Progressive Validation I] Avrim Blum, Adam Kalai, and John Langford Beating the Holdout: Bounds for KFold and Progressive Cross-Validation. COLT99.

[Progressive Validation II] N. Cesa-Bianchi, A. Conconi, and C. Gentile On the generalization ability of on-line learning algorithms IEEE Transactions on Information Theory, 50(9):2050-2057, 2004.

Список литературы

[Importance Aware Updates] Nikos Karampatziakis and John Langford, Importance Weight Aware Gradient Updates UAI 2010.

[Online Convex Programming] Martin Zinkevich, Online convex programming and generalized infinitesimal gradient ascent, ICML 2003.

Список литературы

[Adaptive Updates I] John Duchi, Elad Hazan, and Yoram Singer, Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, COLT 2010 & JMLR 2011.
[Adaptive Updates II] H. Brendan McMahan, Matthew Streeter, Adaptive Bound Optimization for Online Convex Optimization, COLT 2010.
[Scale invariant updates] S.Ross, P.Mineiro, J.Langford, Normalized Online Learning, http://arxiv.org/abs/1305.6646.
[Online BS] Qin, Z., Petricek, V., & Karampatziakis, N. (2013). Efficient Online Bootstrapping for Large Scale Learning. arXiv Preprint `http://arxiv.org/abs/1312.5021`