Метрики и функции потерь

Доп. главы машинного обучения

Андрей Белов

22 сентября 2017

Московский физико-технический институт. ФИВТ. ABBYY

Преподаватели

- Константин Зуев
- Алексей Журавлев
- Андрей Белов
- Алексей Романов

План курса

1. Метрики и функции потерь
2. Методы оптимизации
3. HMM, MEMS, CRF
4. Эмбединги
5. Архитектуры CNN для изображений
6. Архитектуры RNN и CNN для последовательностей
7. Обзор фреймворков

Примеры задач:

8. Классификация текста
9. Классификация изображений
10. Разметка текста
11. Сегментация изображений
12. Object detection
13. Sequence to sequence

План лекции

1. Обзор задач машинного обучения
2. Регрессия
3. Классификация
4. Обучение представления

Обзор задач машинного обучения

Что мы хотим от машинного обучения?

Обычно у нас есть данные $X = \{x_i\}$, $Y = \{y_i\}$

Мы хотим обучить функцию $f(x, \boldsymbol{\theta})$

Для этого мы подбираем параметры θ на обучающей выборке так, чтобы улучшить наши показатели

$$J(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\sim\hat{p}_{data}} L(f(x, \boldsymbol{\theta}), y)$$

То, что нас на самом деле интересует - поведение модели на всех данных

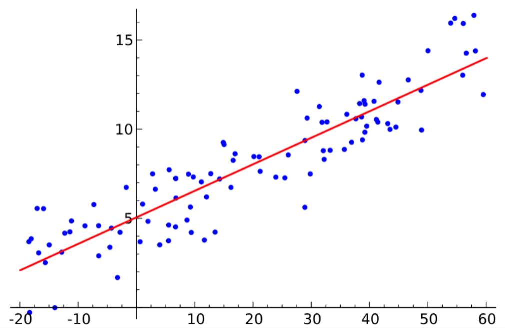
$$J(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\sim p_{data}} L(f(x, \boldsymbol{\theta}), y)$$

Регрессия

Регрессия

$\mathbb{X}$ - пространство объектов

$\mathbb{Y} = \mathbb{R}$ - пространство ответов



Метрики и функции потерь

Метрика - среднее значение ошибки на объекте

$$MeanError(f, X) = \frac{1}{l}\sum_{i=1}^{l} L(y_i, f(x_i))$$

Square loss - квадратичная функция потерь. Метрика - **MSE**

$$L(y, f(x)) = (y - f(x))^2$$

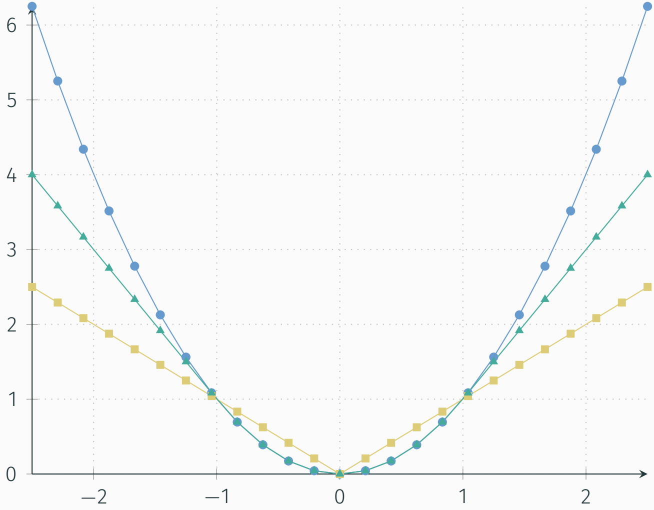
Linear loss - линейная функция потерь. Метрика - **MAE**

$$L(y, f(x)) = |y - f(x)|$$

Huber loss - квадратичная, но линейная для выбросов

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{if } |y - f(x)| < \delta \\ \delta|y - f(x)| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$$

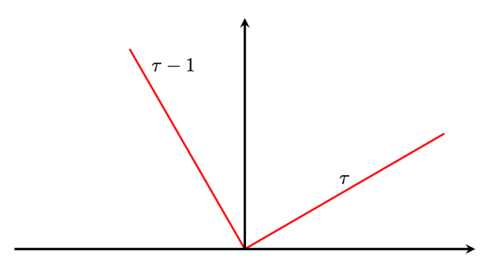
Графики функций потерь



Несимметричные потери

Квантильная ошибка

$$\rho_\tau(f, X) = \frac{1}{l}\sum_{i=1}^{l}((\tau - 1)[y_i < f(x_i)] + \tau[y_i \geq f(x_i)])(y_i - f(x_i))$$



Коэффициент детерминации

$$R^2(f, X) = 1 - \frac{\sum_{i=1}^{l}(y_i - f(x_i))^2}{\sum_{i=1}^{l}(y_i - \bar{y})^2}, \quad \bar{y} = \frac{1}{l}\sum_{i=1}^{l} y_i$$

Показывает, какую долю дисперсии (разнообразия ответов) во всем целевом векторе y модель смогла объяснить.

$R^2 = 1$ - идеальная модель
$R^2 = 0$ - константная модель, возвращающая среднее значение
$R^2 < 0$ - модель хуже константной

Классификация

Классификация

$$\mathbb{Y} = \{+1, -1\}$$

Верность

Верность (accuracy) - доля правильных ответов, которые дает модель.

$$Accuracy = \frac{\sum_i [\hat{y}_i = y_i]}{|Y|}$$

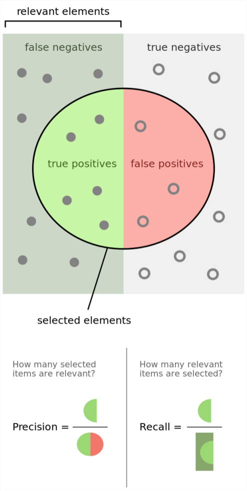
Мало информативна, когда классы не сбалансированы.

Если 90% объектов относятся к положительному классу, а 10% - к отрицательному, то классификатор, всегда выдающий положительный ответ будет работать с верностью *0,9*.

В этом случае можно использовать метрику R^2 на вероятностях, которые вернул класс.

Точность и полнота

	$y=1$	$y=-1$
$a_1(x)=1$	80	20
$a_1(x)=-1$	20	80



Точность

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Полнота

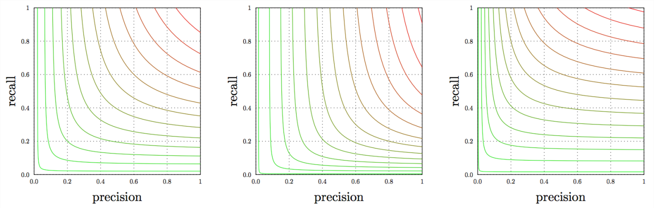
$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

F-мера

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$F_\beta = (1+\beta^2)\frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$$

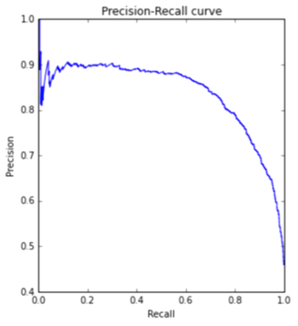
F1-мера



Линии уровня F_1, $F_{0,5}$ и F_2.

Кривая Precision-Recall

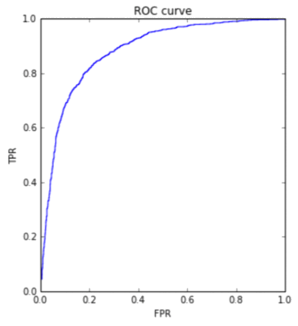
b(x)	0.14	0.23	0.39	0.54	0.73	0.90
y	0	1	0	0	1	1



Площадь под кривой - AUC-PRC.

ROC-кривая

$$FPR = \frac{FP}{FP + TN} \quad TPR = \frac{TP}{TP + FN}$$



Площадь под кривой - ROC-AUC.

Функции потерь для классификации

Идеальная функция потерь - **0-1 loss**

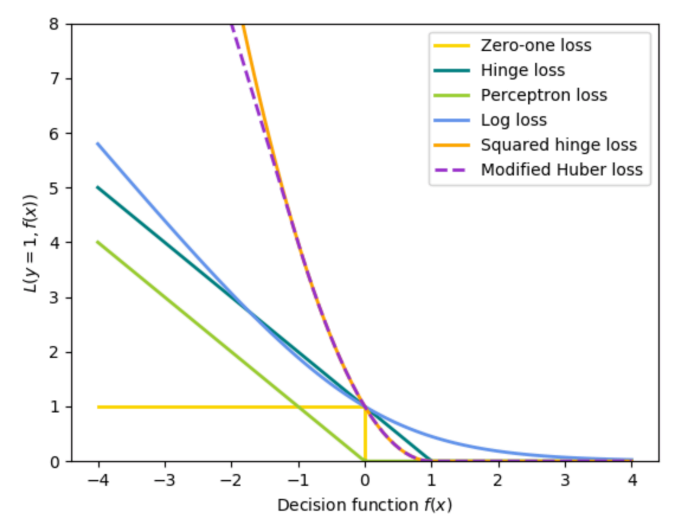
$$L(y, f(x)) = [y \neq f(x)]$$

Но она недифференцируемая и разрывная, поэтому плохо подходит для оптимизации.

Обычно используют суррогатные функции потерь, которые оценивают сверху 0-1 loss.

Функции потерь для классификации

$$f(x) \in R, \quad y \in \{-1, 1\} \qquad M = yf(x)$$



Функции потерь для классификации

Logistic loss - логистическая функция потерь

$$L(y, f(x)) = \frac{1}{\ln 2} \ln(1 + e^{-yf(x)})$$

- Более устойчива к выбросам, чем квадратичная функция потерь: при больших отрицательных аргументах становится похожа на линейную
- Штраф > 0 для всех точек. Это позволяет получать более уверенные модели.
- Но может мешать: за штрафами большого количества уже правильных ответов может потеряется штраф на настоящих ошибках.
  Особенно актуально в случае несбалансированных классов.

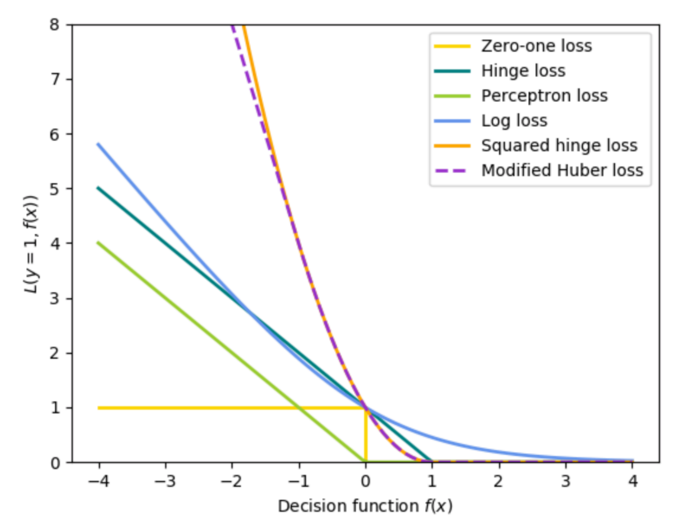
Hinge loss

$$L(y, f(x)) = max(0, 1 - yf(x))$$

- Устойчива к выбросам, т.к. линейно зависит от величины f(x) на ошибках.
- Штраф = 0 для объектов с f(x) > 1. Обучение работает только над ошибками: объекты с f(x) > 1 не влияют на градиент.
  (эта идея подходит и для регрессии)
- Модель стремится назначить одному классу скор f(x) > 1, а другому < 1. То есть, сделать зазор между классами. Это положительно сказывается на устойчивости решения (обобщающей способности).

Функции потерь для классификации

$$f(x) \in R, \quad y \in \{-1, 1\} \qquad M = yf(x)$$



Cross-entropy loss

Здесь $y_i \in \{1, \ldots, N\}$, где N - число классов. $f(x_i) \in [0, 1]^N$

Кросс-энтропия между "правильным" распределением p и оцениваемым q определяется как:

$$H(p, q) = -\sum_{x} p(x) \log q(x)$$

Пусть $p(x_i) = [0, 0, \ldots 1, \ldots 0, 0]$ содержит 1 в элементе y_i.
А $q(x_i)$ - содержит вероятности отнесения x_i к соответствующему классу.

Тогда, **cross-entropy loss** записывается как

$$L(y, f(x)) = -\sum_{k=1}^{N} p_i(x) \log f_i(x) = -\sum_{k=1}^{N} [y = k] \log f_k(x) = -\log f_{y_i}(x)$$

Кросс-энтропия минимальна, если $p = q$.
Поэтому наша цель - минимизировать её.

Связь cross entropy с максимизацией правдоподобия

Если $f_k(x)$ - вероятность, с которой классификатор относит x к классу k, то правдоподобие выборки $(x_i, y_i)_{i=1}^{l}$ записывается как

$$\mathcal{L} = \prod_{i=1}^{l} f_{y_i}(x_i)$$

Чтобы получить наилучшую модель, мы максимизируем правдоподобие $\mathcal{L}$. Что эквивалентно минимизации $-\log \mathcal{L}$.

$$-\log \mathcal{L} = -\log \left(\prod_{i=1}^{l} f_{y_i}(x_i) \right) = -\sum_{i=1}^{l} \log f_{y_i}(x_i)$$

Что равно среднему значению cross entropy loss на объектах выборки.

Связь cross entropy с логистической функцией потерь

Ранее мы рассматривали функции $f(x) \in \mathbb{R}$ и $y_i \in \{0, 1\}$.

Например, f(x) для линейного классификатора $f(x) = Wx + b$.

Как применить cross entropy loss для такой $f(x)$?

Привести к диапазону $[0, 1]$ с помощью сигмоиды $\sigma = \frac{1}{1+e^{-x}}$. Тогда вероятность отнесения x к классу $y = 1$ равна

$$g(x) = \sigma(f(x)) = \frac{1}{1 + e^{-f(x)}}$$

А к классу $y_i \in \{-1, 1\}$:

$$g(x, y_i) = \frac{1}{1 + e^{-y_i f(x)}}$$

Подставив это значение в формулу cross entropy loss, получим логистическую функцию потерь.

Softmax

В случае многоклассовой классификации часто мы имеем функцию $f(x) \in \mathbb{R}^N \quad y_i \in \{0, 1\}$

Чтобы получить вероятностное распределение по классам, применяем softmax.

$$g_i(f(x)) = \frac{e^{f_i(x)}}{\sum_{j=1}^{N} e^{f_j}}$$

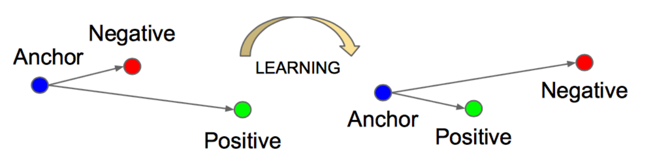
Обучение представления

Contrastive loss

$$L_{cts} = \sum_{i,j}^{N} [y_{ij} d + (1 - y_{ij}) max(0, \alpha_{cts} - d)]$$

$$d = \|f(x_i) - f(x_j)\|_2^2$$

Triplet loss

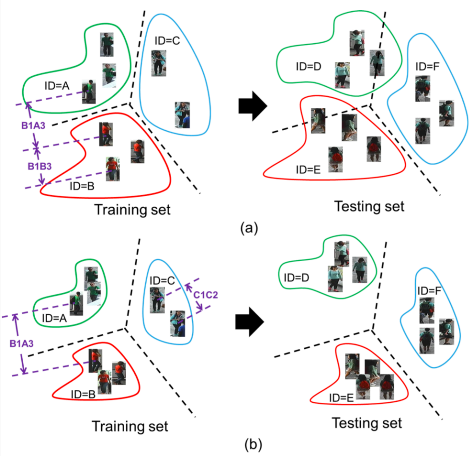


$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2 \ , \quad (1)$$

$$\forall \left(f(x_i^a), f(x_i^p), f(x_i^n)\right) \in \mathcal{T} \ . \quad (2)$$

$$\sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

Quadruplet loss



$$L_{quad} = \sum_{i,j,k}^{N} [g(x_i, x_j)^2 - g(x_i, x_k)^2 + \alpha_1]_+$$

Ваши вопросы