Машинное обучение

Лекция 5

Решающие деревья и ансамбли деревьев: дополнительные темы

Виктор Кантор

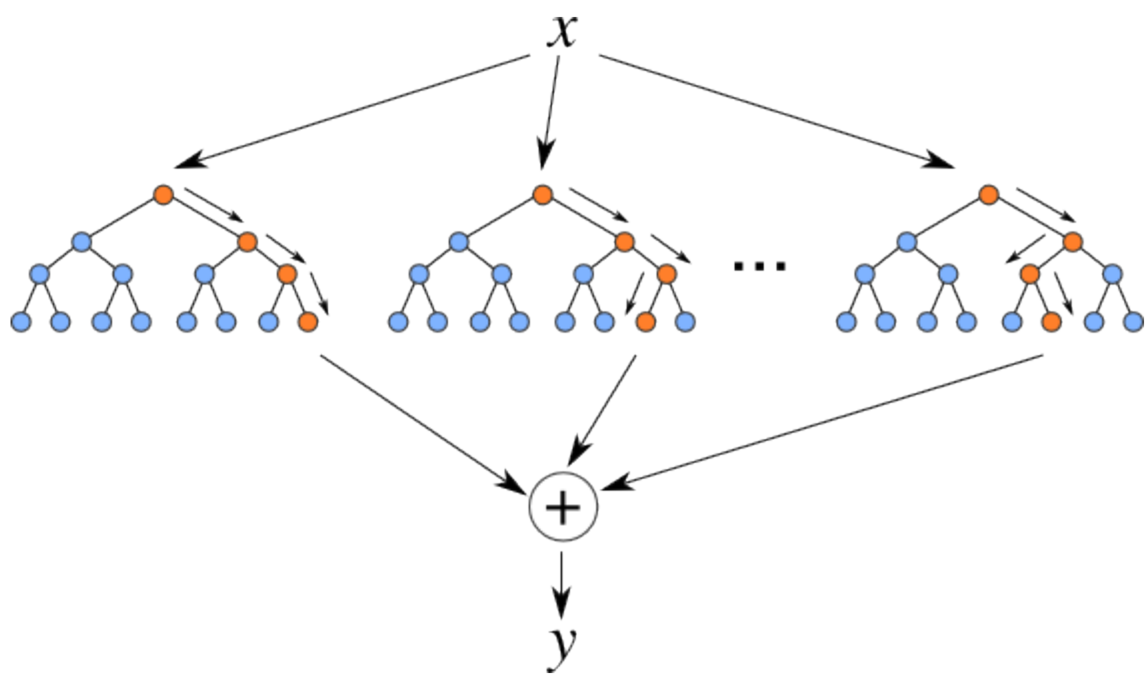
План

I. Ансамбли решающих деревьев
   a) Анализ RF и GBDT
   b) XGBoost

II. Решающие деревья
   a) Критерии информативности
   b) Прунниинг
   c) Категориальные признаки
   d) Пропущенные значения
   e) ID3, C4.5, CART

I. Ансамбли решающих деревьев: дополнительные темы

Random Forest



1. Бэггинг над деревьями
2. Рандомизированные разбиения в деревьях: выбираем k случайных признаков и ищем наиболее информативное разбиение по ним

Ошибка усредненной модели

Вопрос:

Как можно оценить (сверху и снизу) матожидание квадрата отклонения прогноза усредненной модели от ответа?

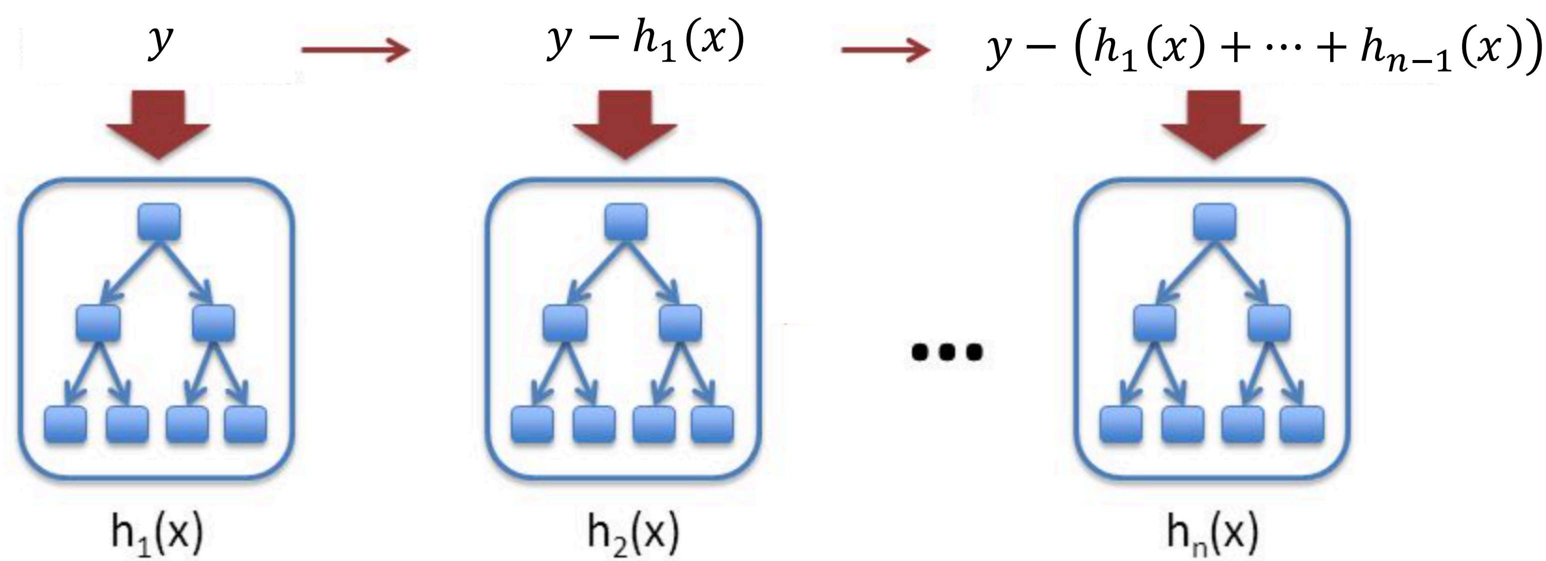
Out-of-bag оценка

При бутстрепе часть выборки не используется для построения дерева, значит ее можно использовать для контроля:

$$\mathrm{OOB} = \sum_{i=1}^{\ell} L\left(y_i, \frac{1}{\sum_{n=1}^{N}[x_i \notin X_n^\ell]} \sum_{n=1}^{N} [x_i \notin X_n^\ell] b_n(x_i)\right)$$

Идея Gradient Boosted Decision Trees

$$a_n(x) = h_1(x) + \cdots + h_n(x)$$



GBM в наиболее общем виде

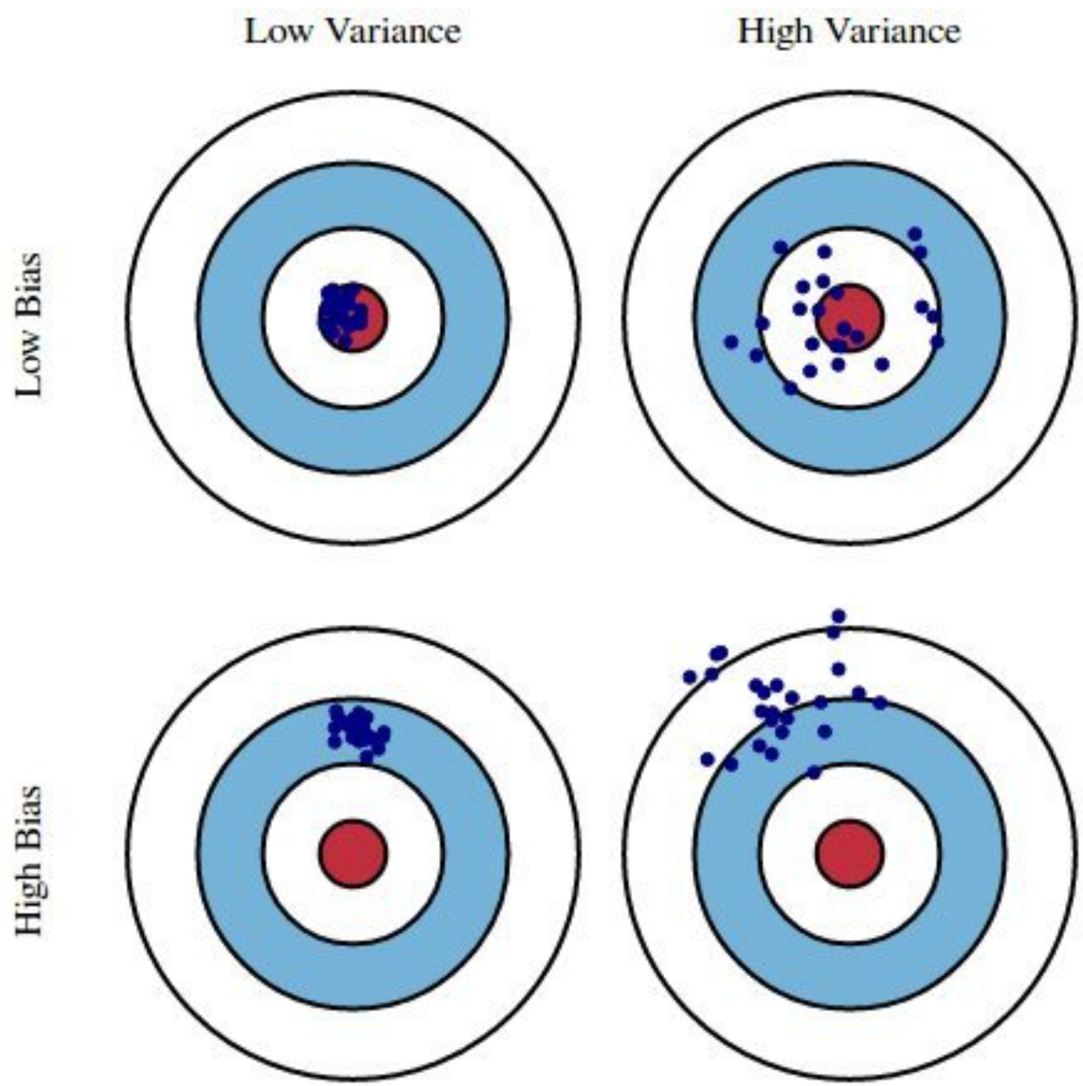
1. Обучаем первый базовый алгоритм $h_1, \ \beta_1 = 1$
2. Повторяем в цикле по t от 2 до T:

$$h_t = \underset{h}{\operatorname{argmin}} \sum_{i=1}^{l} \tilde{L}\left(h(x_i), -\frac{\partial L(\hat{y}_i, y_i)}{\partial \hat{y}_i}\right)$$

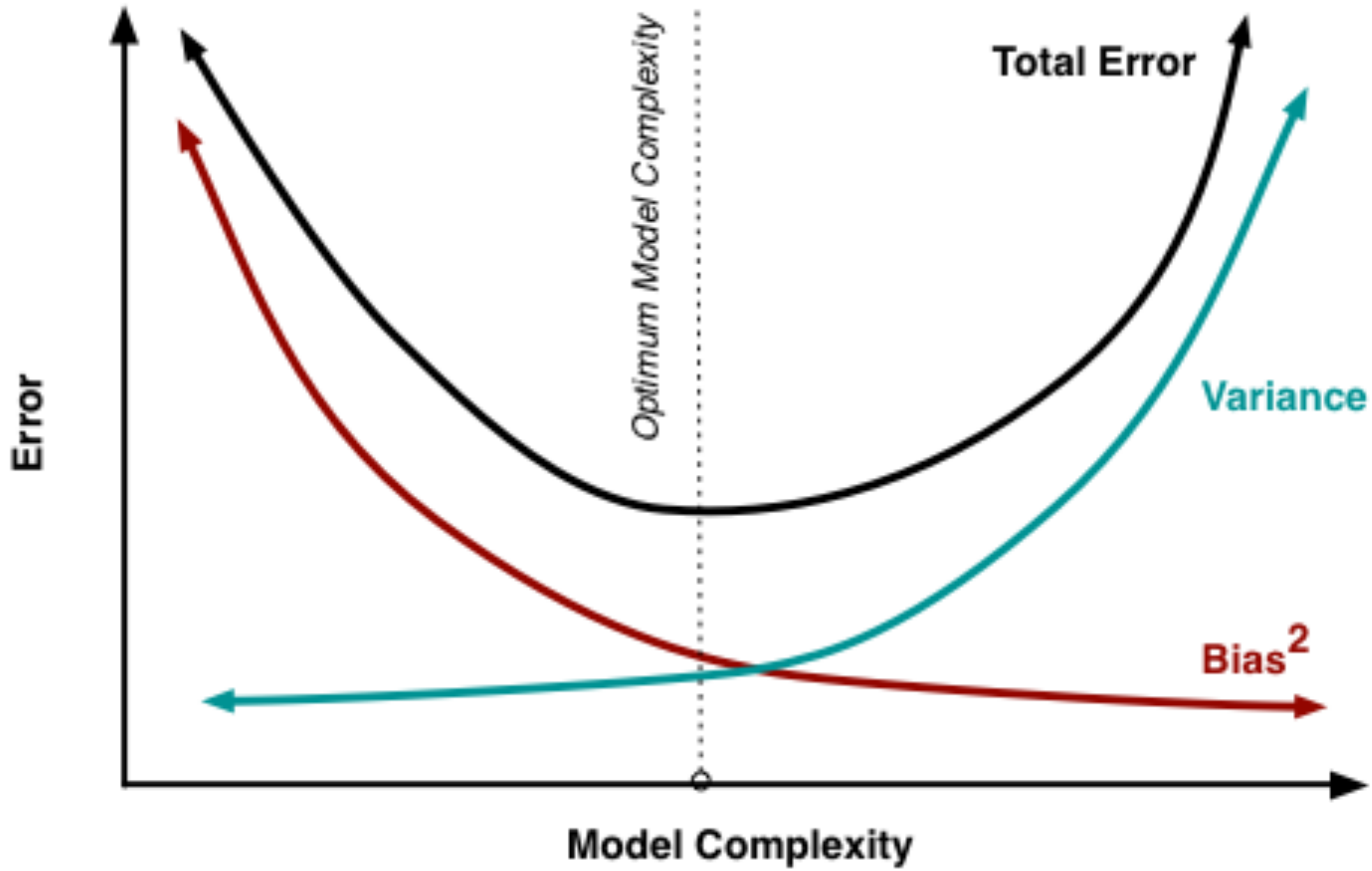
выбираем β_t

Здесь $Q(\hat{y}, y) = \sum_{i=1}^{l} L(\hat{y}_i, y_i)$ $\qquad \hat{y}_i = a_{t-1}(x_i)$

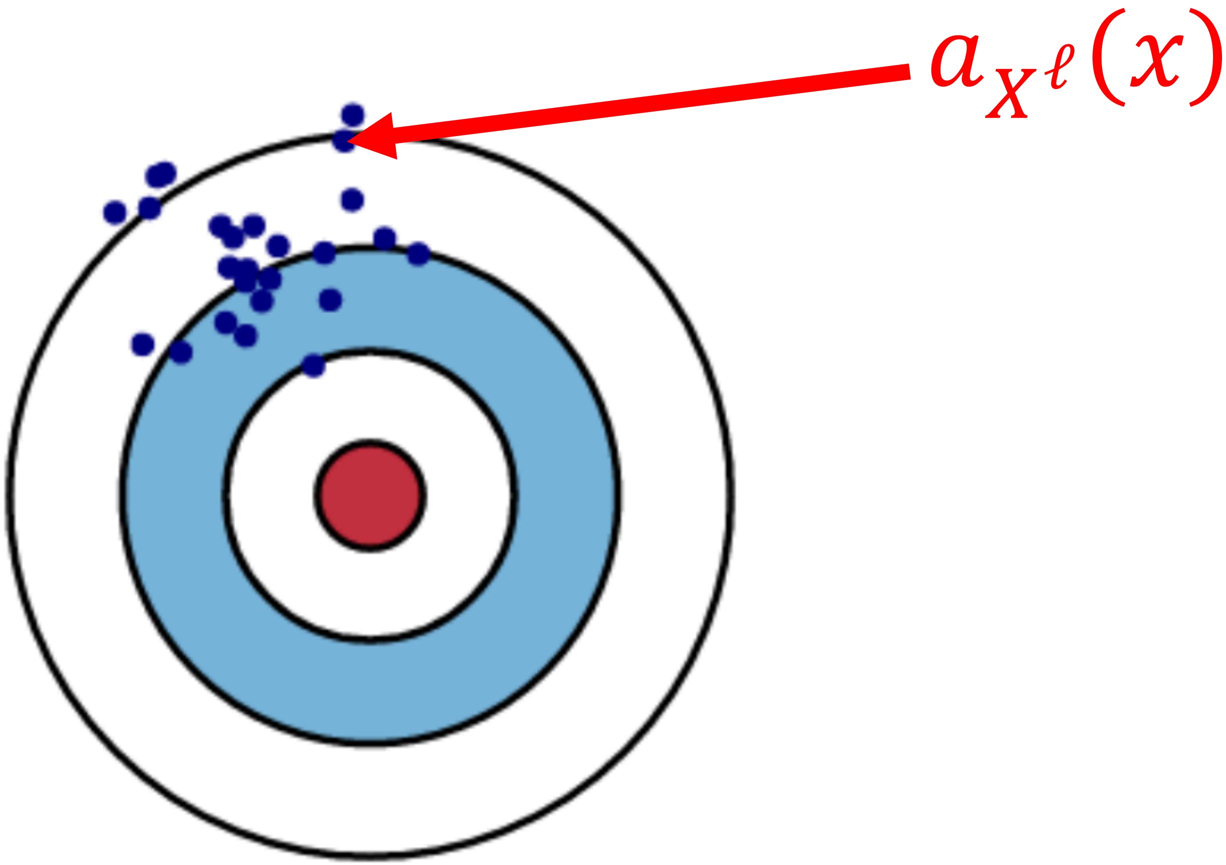
Bias-variance trade-off



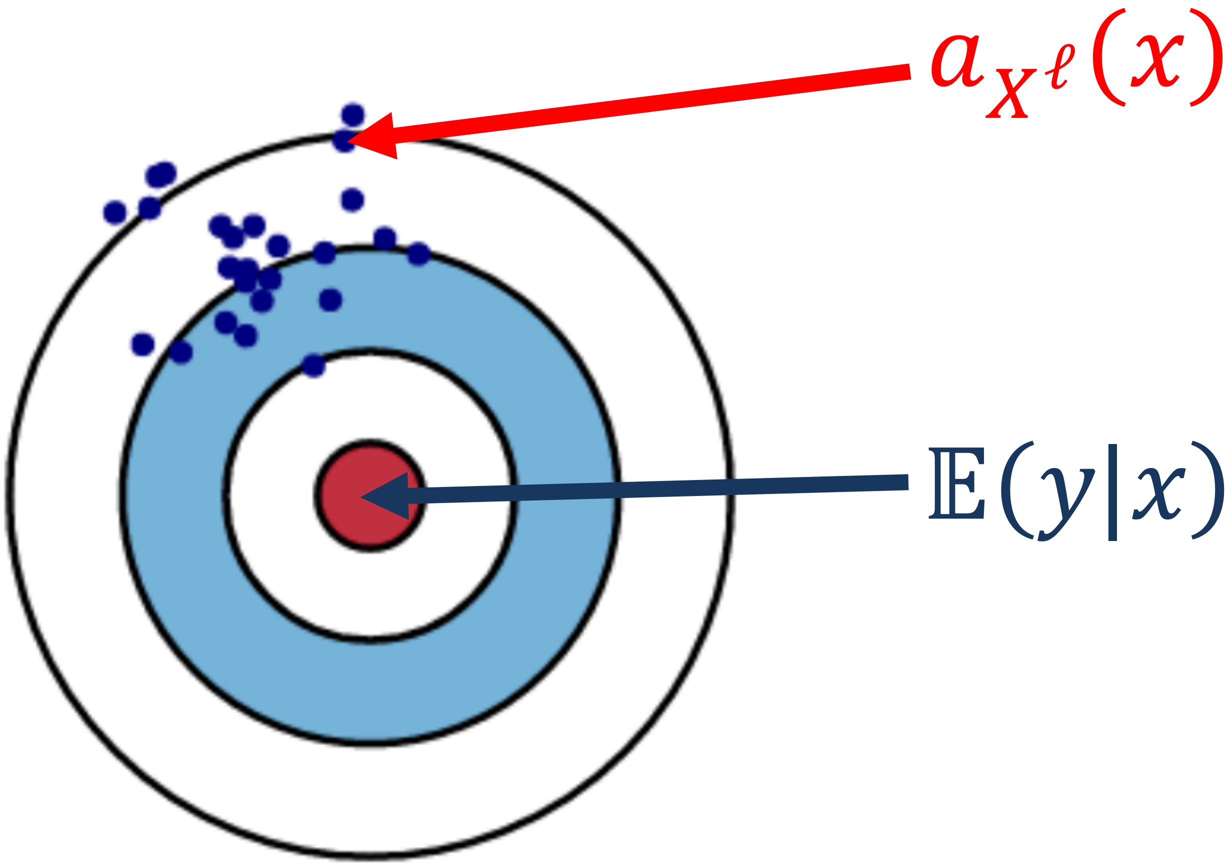
Недообучение и переобучение



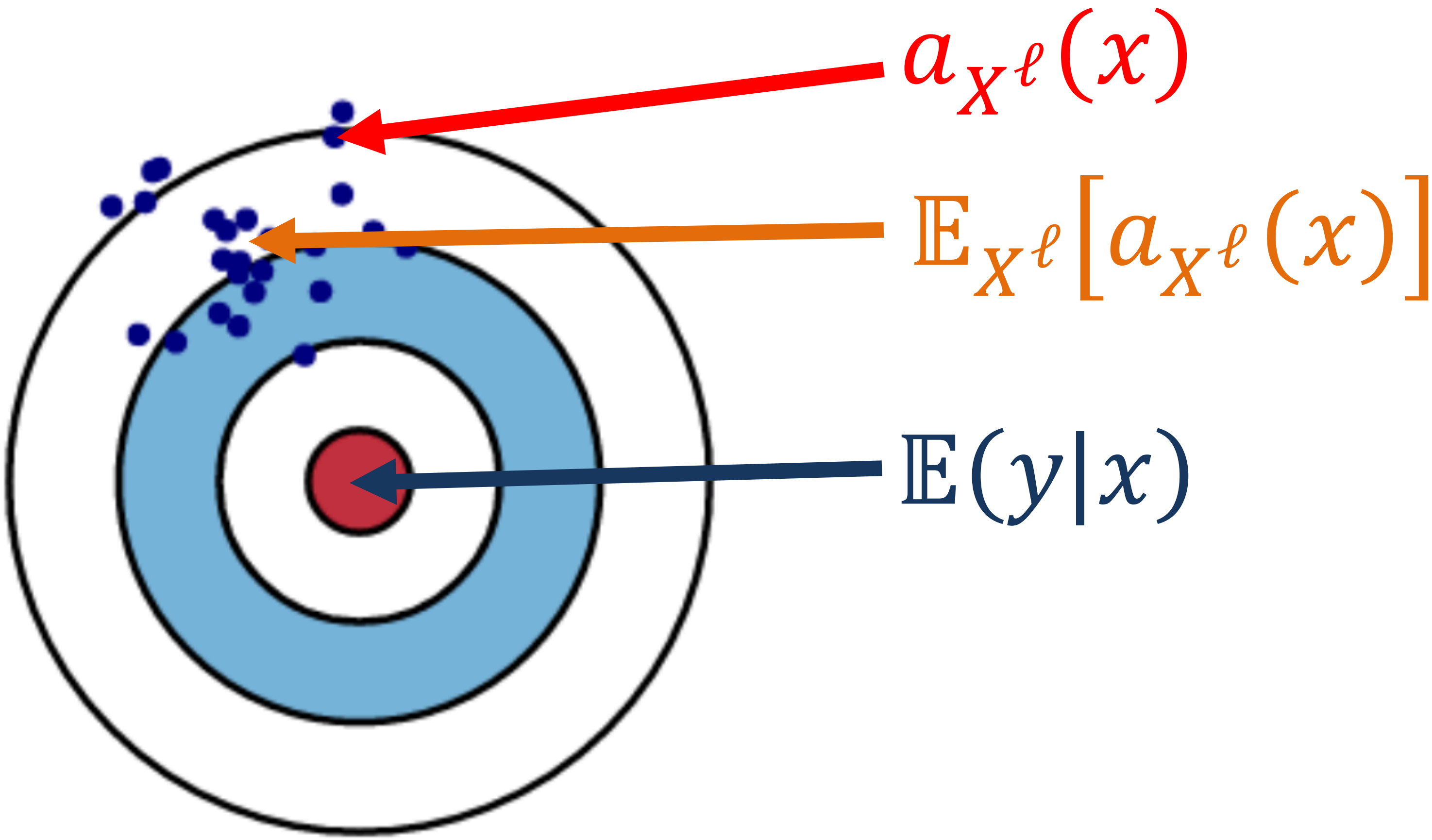
Смещение и разброс



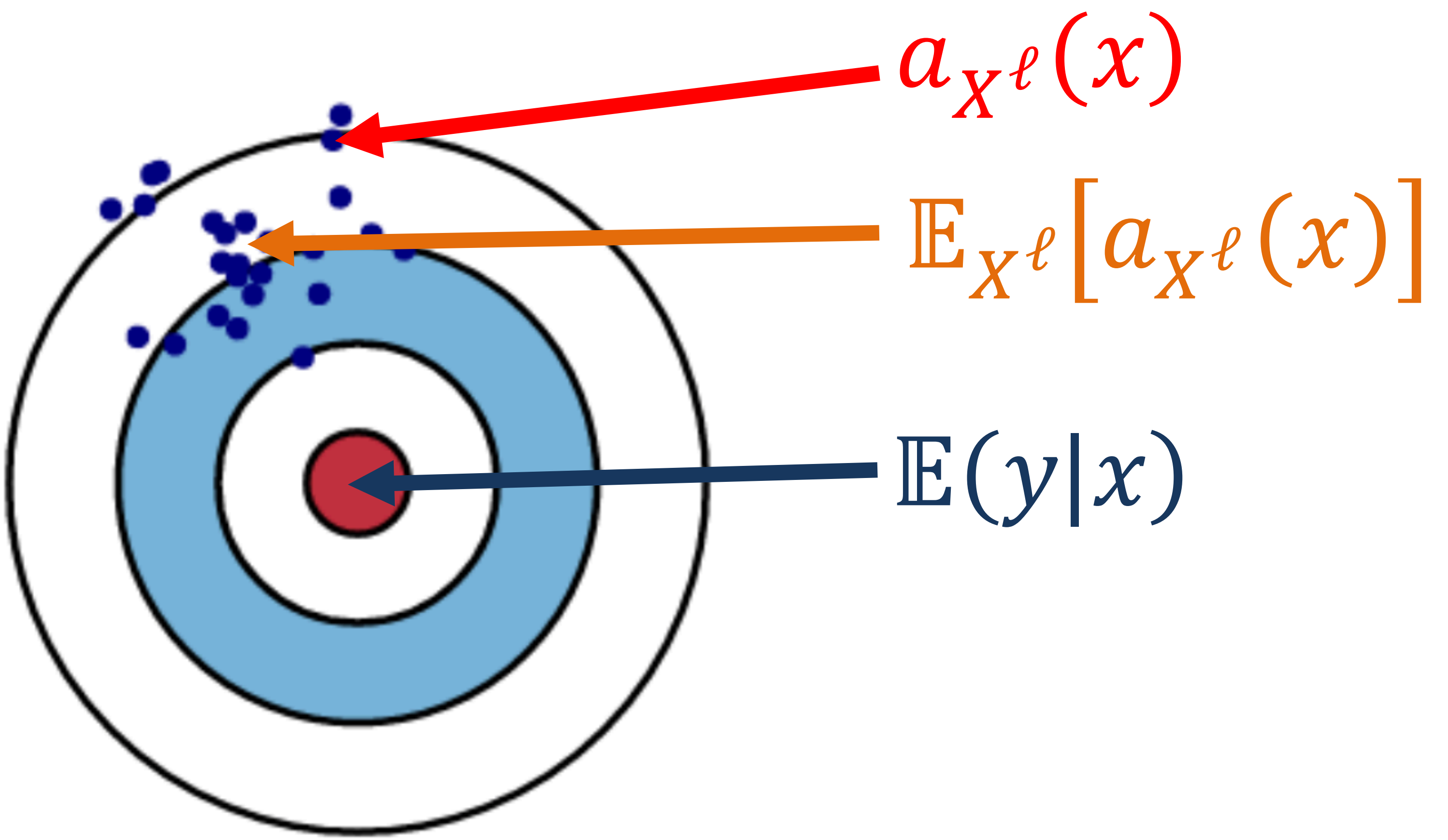
Смещение и разброс



Смещение и разброс

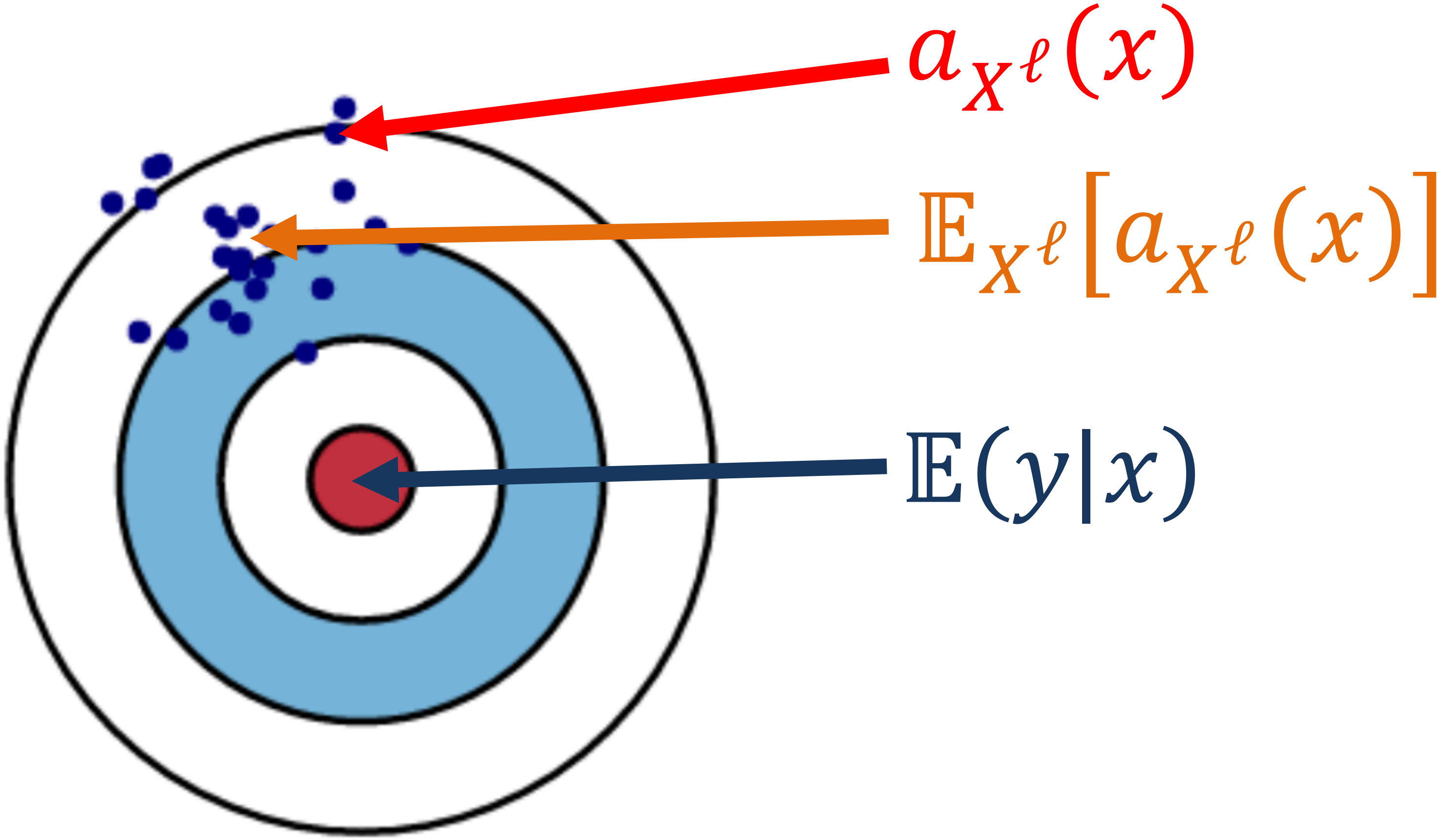


Смещение и разброс



$$Bias^2(a) = \mathbb{E}_{x,y}\left(\mathbb{E}(y|x) - \mathbb{E}_{X^\ell}[a_{X^\ell}(x)]\right)^2$$

Смещение и разброс



$$Var(a) = \mathbb{E}_{x,y}\mathbb{E}_{X^\ell}\left(a_{X^\ell}(x) - \mathbb{E}_{X^\ell}\left[a_{X^\ell}(x)\right]\right)^2$$

Bias-variance-noise decomposition

$$\mathbb{E}_{x,y}\mathbb{E}_{X^\ell}\left(y - a_{X^\ell}(x)\right)^2 =$$

$$\underbrace{\mathbb{E}_{x,y}\big(y - \mathbb{E}(y|x)\big)^2}_{Noise^2} +$$

$$+ \underbrace{\mathbb{E}_{x,y}\Big(\mathbb{E}(y|x) - \mathbb{E}_{X^\ell}\big[a_{X^\ell}(x)\big]\Big)^2}_{Bias^2} +$$

$$+ \underbrace{\mathbb{E}_{x,y}\mathbb{E}_{X^\ell}\Big(a_{X^\ell}(x) - \mathbb{E}_{X^\ell}\big[a_{X^\ell}(x)\big]\Big)^2}_{Variance}$$

Bias-variance trade-off и деревья

С ростом количества деревьев:

- В GBM над деревьями – уменьшается смещение
- В Random Forest – уменьшается разброс

Разброс при усреднении моделей

Вопрос:

Как связаны разброс усредненной модели и базовой?

Регуляризация в GBM

- Метод сокращения шага:

$$a_N(x) = a_{N-1}(x) + \eta\beta_N h_N(x)$$

β_N - шаг наискорейшего спуска

$\eta \in (0, 1]$ - темп обучения

Регуляризация в GBM

- Метод сокращения шага:

$$a_N(x) = a_{N-1}(x) + \eta\beta_N h_N(x)$$

β_N - шаг наискорейшего спуска

$\eta \in (0, 1]$ - темп обучения

- Стохастический градиентный бустинг:

приближаем градиент по случайной подвыборке

eXtreme Gradient Boosting (XGBoost)

$$\sum_{i=1}^{\ell} L(y_i, a_{N-1}(x_i) + b(x_i)) \to \min_b$$

$$s = \left(-\left.\frac{\partial L}{\partial z}\right|_{z=a_{N-1}(x_i)} \right)_{i=1}^{\ell} = -\nabla_s \sum_{i=1}^{\ell} L(y_i, a_{N-1}(x_i) + s_i)$$

$$b_N(x) = \arg\min_{b \in \mathcal{A}} \sum_{i=1}^{\ell} (b(x_i) - s_i)^2$$

$$\Downarrow$$

$$b_N(x) = \arg\min_{b \in \mathcal{A}} \sum_{i=1}^{\ell} \left(b(x_i) - \frac{s_i}{h_i} \right)^2 \qquad h_i = \left.\frac{\partial^2 L}{\partial z^2}\right|_{z=a_{N-1}(x_i)}$$

eXtreme Gradient Boosting (XGBoost)

$$b_N(x) = \arg\min_{b \in \mathcal{A}} \sum_{i=1}^{\ell} \left(b(x_i) - \frac{s_i}{h_i} \right)^2$$

$$b(x) = \sum_{j=1}^{J} b_j [x \in R_j]$$

$$\sum_{i=1}^{\ell} \left(-s_i b(x_i) + \frac{1}{2} h_i b^2(x_i) \right) + \lambda J + \frac{\mu}{2} \sum_{j=1}^{J} b_j^2 \to \min_b$$

$$\sum_{j=1}^{J} \left\{ \underbrace{\left(-\sum_{i \in R_j} s_i \right)}_{=-S_j} b_j + \frac{1}{2} \left(\mu + \underbrace{\sum_{i \in R_j} h_i}_{=H_j} \right) b_j^2 + \lambda \right\}$$

eXtreme Gradient Boosting (XGBoost)

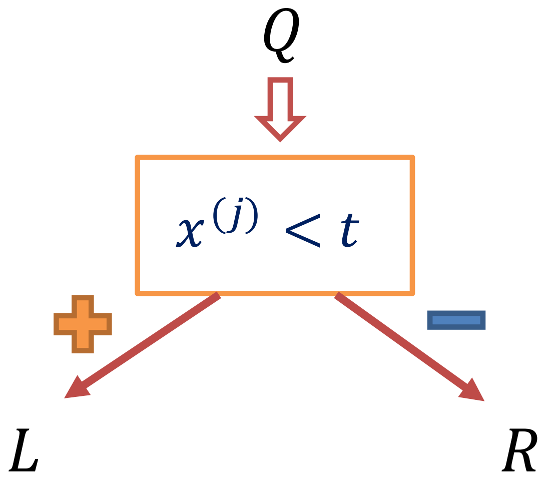
$$b_j = \frac{S_i}{H_j + \mu}$$

$$H(b) = \frac{1}{2} \sum_{j=1}^{J} \frac{S_j^2}{H_j + \mu} + \lambda J$$

$$H(b_l) + H(b_r) - H(b) - \lambda \to \max$$

II. Решающие деревья: дополнительные темы

Выбор разбиения



$$G(j,t) = \frac{|L|}{|Q|} H(L) + \frac{|R|}{|Q|} H(R) \to \min_{j,t}$$

Критерии построения разбиений

$H(R) -$ мера «неоднородности» множества R

Критерии построения разбиений

$H(R)$ – мера «неоднородности» множества R

Пусть мы решаем задачу классификации на 2 класса,
p_0, p_1 – доли объектов классов 0 и 1 в R

1) Misclassification criteria: $H(R) = 1 - \max\{p_0, p_1\}$
2) Entropy criteria: $H(R) = -p_0 \ln p_0 - p_1 \ln p_1$
3) Gini criteria: $H(R) = 1 - {p_0}^2 - {p_1}^2 = 2p_0p_1$

Критерии построения разбиений

$H(R)$ – мера «неоднородности» множества R

Пусть мы решаем задачу классификации на K классов,
$p_1, \ldots, p_K$ – доли объектов классов 1, …, K в R

1) Misclassification criteria: $H(R) = 1 - p_{max}$
2) Entropy criteria:
$$H(R) = -\sum_{k=1}^{K} p_k \ln p_k$$
3) Gini criteria:
$$H(R) = \sum_{k=1}^{K} p_k (1 - p_k)$$

Критерии построения разбиений

$H(R)$ – мера «неоднородности» множества R

Чтобы решать задачу регрессии, достаточно взять среднеквадратичную ошибку в качестве $H(R)$:

$$H(R) = \frac{1}{|R|} \sum_{x_i \in R} (y_i - \bar{y})^2$$

Критерии построения разбиений

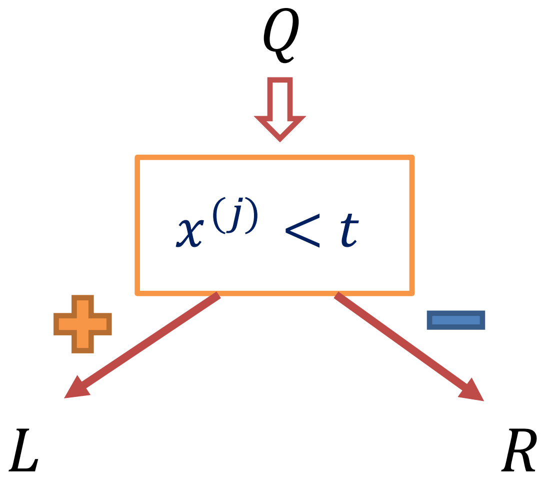
$H(R)$ – мера «неоднородности» множества R

Чтобы решать задачу регрессии, достаточно взять среднеквадратичную ошибку в качестве $H(R)$:

$$H(R) = \frac{1}{|R|} \sum_{x_i \in R} (y_i - \bar{y})^2$$

$$\bar{y} = \frac{1}{|R|} \sum_{x_i \in R} y_i$$

Критерии информативности



$$I(Q, j, t) = H(Q) - \frac{|L|}{|Q|} H(L) - \frac{|R|}{|Q|} H(R)$$

Gini

$$I(Q,j,t) = H(Q) - \frac{|L|}{|Q|}H(L) - \frac{|R|}{|Q|}H(R)$$

$$H(R) = \sum_{k=1}^{K} p_k(1 - p_k)$$

Information gain

$$I(Q,j,t) = H(Q) - \frac{|L|}{|Q|}H(L) - \frac{|R|}{|Q|}H(R)$$

$$H(R) = -\sum_{k=1}^{K} p_k \ln p_k$$

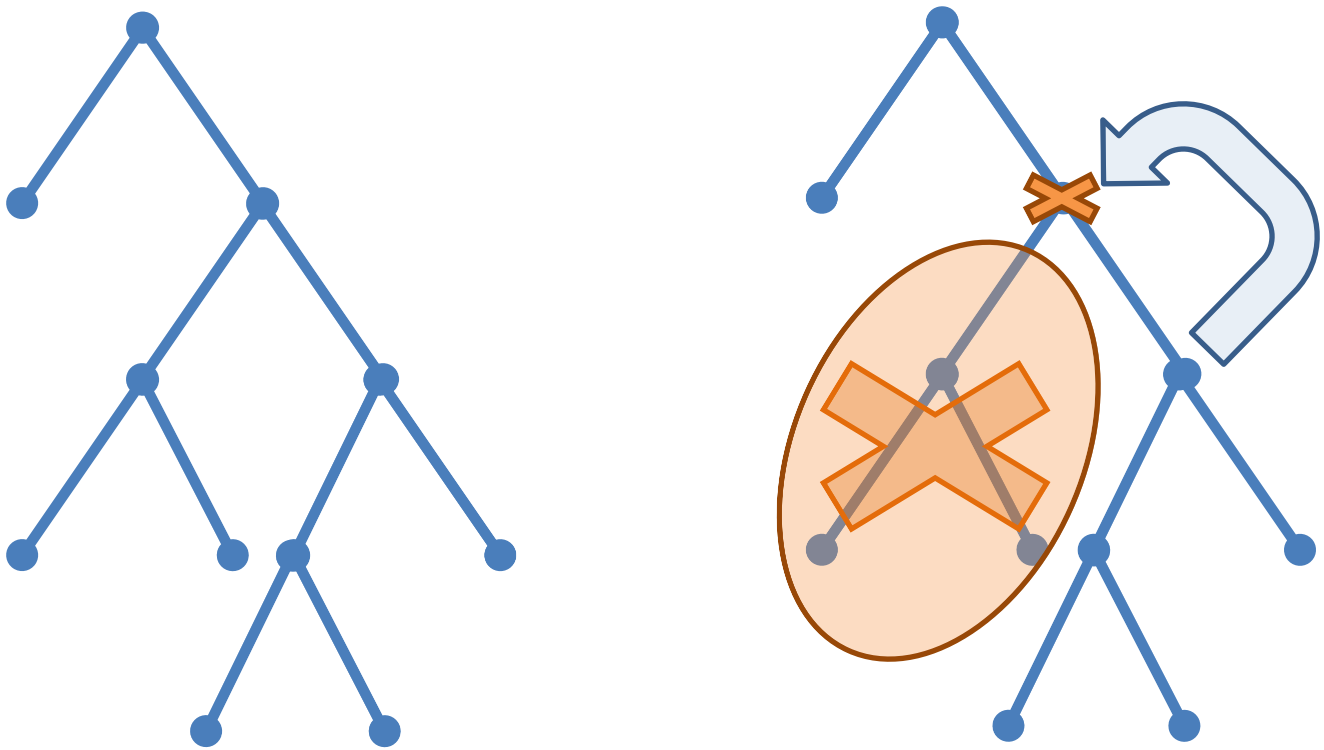
Чем полезен Information Gain

- Information Gain имеет непосредственное отношение к теории информации
- На основе утверждений из теории информации можно получить оценку максимальной точности при заданной полноте и наоборот
- Подробней см. в книге «Теория информации и распознавание образов» Щепина

Prunning

- Pre-prunning:
  - Ограничиваем рост дерева до того как оно построено
  - Если в какой-то момент информативность признаков в разбиении меньше порога – не разбиваем вершину
- Post-prunning:
  - Упрощаем дерево после того как дерево построено

Post-prunning



Cost-complexity prunning

$$C_\alpha(T) = R(T) + \alpha|T| \rightarrow min$$

Cost-complexity prunning

$$C_\alpha(T) = R(T) + \alpha|T| \to min$$

$$T_0 \succ T_1 \succ \cdots \succ T_K$$

$$0 < \alpha_0 < \alpha_1 < \cdots < \alpha_K < \infty$$

$$T_i = \underset{\alpha}{\operatorname{argmin}}\, C_\alpha(T) \qquad \forall \alpha \in [\alpha_i, \alpha_{i+1})$$

Категориальные признаки

$$x^{(j)} \in Q = \{u_1, \dots, u_q\}$$

$$Q = Q_1 \sqcup Q_2 \qquad\qquad b(x) = \left[x^{(j)} \in Q_1\right]$$

Всего нужно перебрать разбиений: $\quad 2^{q-1} - 1$

Категориальные признаки

$R(u) -$ множество объектов, попавших в текущую вершину и имеющих $x^{(j)} = u$

Упорядочим $u_1, \dots, u_q$ и получим $u_{(1)}, \dots, u_{(q)}$:

$$\frac{1}{\left|R\left(u_{(1)}\right)\right|} \sum_{x_i \in R\left(u_{(1)}\right)} [y_i = +1] < \dots < \frac{1}{\left|R\left(u_{(q)}\right)\right|} \sum_{x_i \in R\left(u_{(q)}\right)} [y_i = +1]$$

Категориальные признаки

$R(u) -$ множество объектов, попавших в текущую вершину и имеющих $x^{(j)} = u$

Упорядочим $u_1, \dots, u_q$ и получим $u_{(1)}, \dots, u_{(q)}$:

$$\frac{1}{|R(u_{(1)})|} \sum_{x_i \in R(u_{(1)})} [y_i = +1] < \dots < \frac{1}{|R(u_{(q)})|} \sum_{x_i \in R(u_{(q)})} [y_i = +1]$$

Закодируем $u_{(k)} \longmapsto k$ и будем работать как с вещественным признаком

Категориальные признаки

$R(u)$ − множество объектов, попавших в текущую вершину и имеющих $x^{(j)} = u$

Упорядочим $u_1, \dots, u_q$ и получим $u_{(1)}, \dots, u_{(q)}$:

$$\frac{1}{\left|R\left(u_{(1)}\right)\right|} \sum_{x_i \in R\left(u_{(1)}\right)} [y_i = +1] < \dots < \frac{1}{\left|R\left(u_{(q)}\right)\right|} \sum_{x_i \in R\left(u_{(q)}\right)} [y_i = +1]$$

Закодируем $u_{(k)} \longmapsto k$ и будем работать как с вещественным признаком

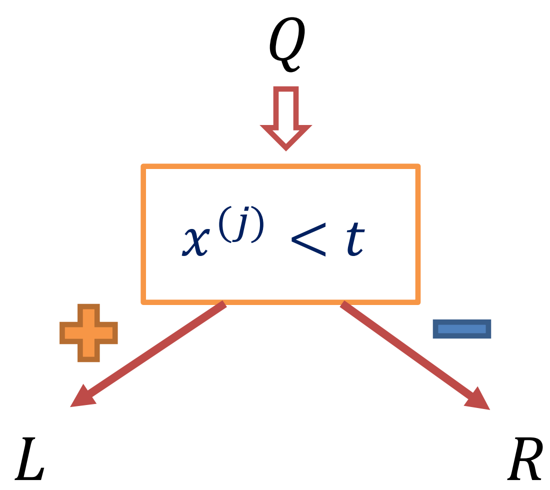
Можно показать: результат для критерия Джини и энтропийного критерия – тот же, как для перебора

Категориальные признаки

В задаче регрессии:

$$\frac{1}{\left|R\left(u_{(1)}\right)\right|} \sum_{x_i \in R\left(u_{(1)}\right)} y_i < \dots < \frac{1}{\left|R\left(u_{(q)}\right)\right|} \sum_{x_i \in R\left(u_{(q)}\right)} y_i$$

Пропущенные значения



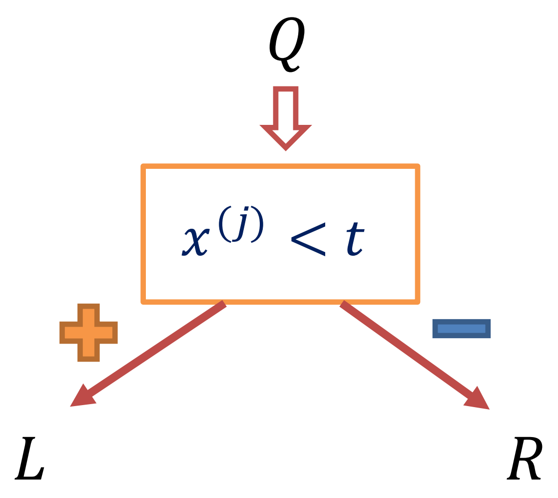
$$G(Q, j, t) = \frac{|L|}{|Q|} H(L) + \frac{|R|}{|Q|} H(R)$$

Пусть $x^{(j)}$ не определен для $V \subset Q$, подправим $G(Q, j, t)$:

$$G(Q, j, t) = \frac{|Q \backslash \mathrm{V}|}{|Q|} G(Q \backslash \mathrm{V}, j, t)$$

Если разбиение по $x^{(j)}$ окажется лучшим, добавим объекты V и в левое, и в правое поддерево

Пропущенные значения



Также можно учитывать объекты из V с весом $\frac{|L|}{|Q|}$ в левом поддереве и $\frac{|R|}{|Q|}$ в правом

При применении также – например, усредняем с этими весами прогноз вероятности класса от левого и правого поддерева

Пропущенные значения

Другой способ обработки: суррогатные предикаты – разбиваем множество V по другому признаку (не пропущенному), разбиение по которому для остальных вершин максимально похоже на наилучшее

ID3: Iterative Dichotomizer 3

- Энтропийный критерий или information gain
- Бинарные признаки, т.е. можно считать, что энтропия считается не для конкретного разбиения, а для признака
- Строим, пока энтропия уменьшается или пока в листе не будет только один класс

C4.5

- Information Gain Ratio
- Добавлены сплиты вещественных признаков по порогу
- Поддержка пропущенных значений: объекты с пропусками просто игнорируются при построении, а потом берутся с весами
- Пост-пруннинг: Error-Based Pruning, удаление вершин на основе оценок обобщающей способности

Information gain ratio

$$I(Q,j,t) = \frac{H(Q) - \frac{|L|}{|Q|}H(L) - \frac{|R|}{|Q|}H(R)}{-\frac{|L|}{|Q|}\ln\frac{|L|}{|Q|} - \frac{|R|}{|Q|}\ln\frac{|R|}{|Q|}}$$

$$H(R) = -\sum_{k=1}^{K} p_k \ln p_k$$

CART

- Привычный нам выбор разбиений
- Могут решать задачу регрессии
- Как правило – критерий Джини для классификации и MSE для регрессии
- Minimal cost-complexity pruning
- Выбор дерева по качеству на тестовой выборке или по V-fold
- Обработка пропусков суррогатными предикатами

CART: построение дерева

$$G(j,t) = \frac{|L|}{|Q|}H(L) + \frac{|R|}{|Q|}H(R) \to \min_{j,t}$$

Для классификации:

$$H(R) = \sum_{k=1}^{K} p_k(1 - p_k)$$

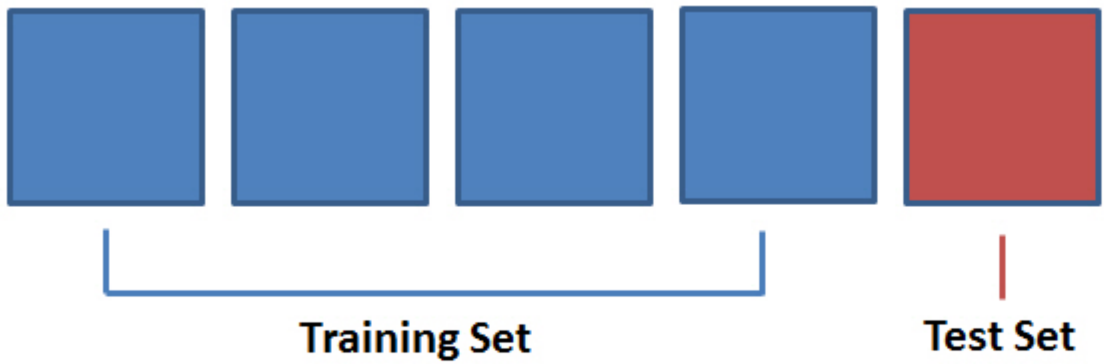
Для регрессии:

$$H(R) = \frac{1}{|R|}\sum_{x_i \in R}(y_i - \bar{y})^2$$

$$\bar{y} = \frac{1}{|R|}\sum_{x_i \in R} y_i$$

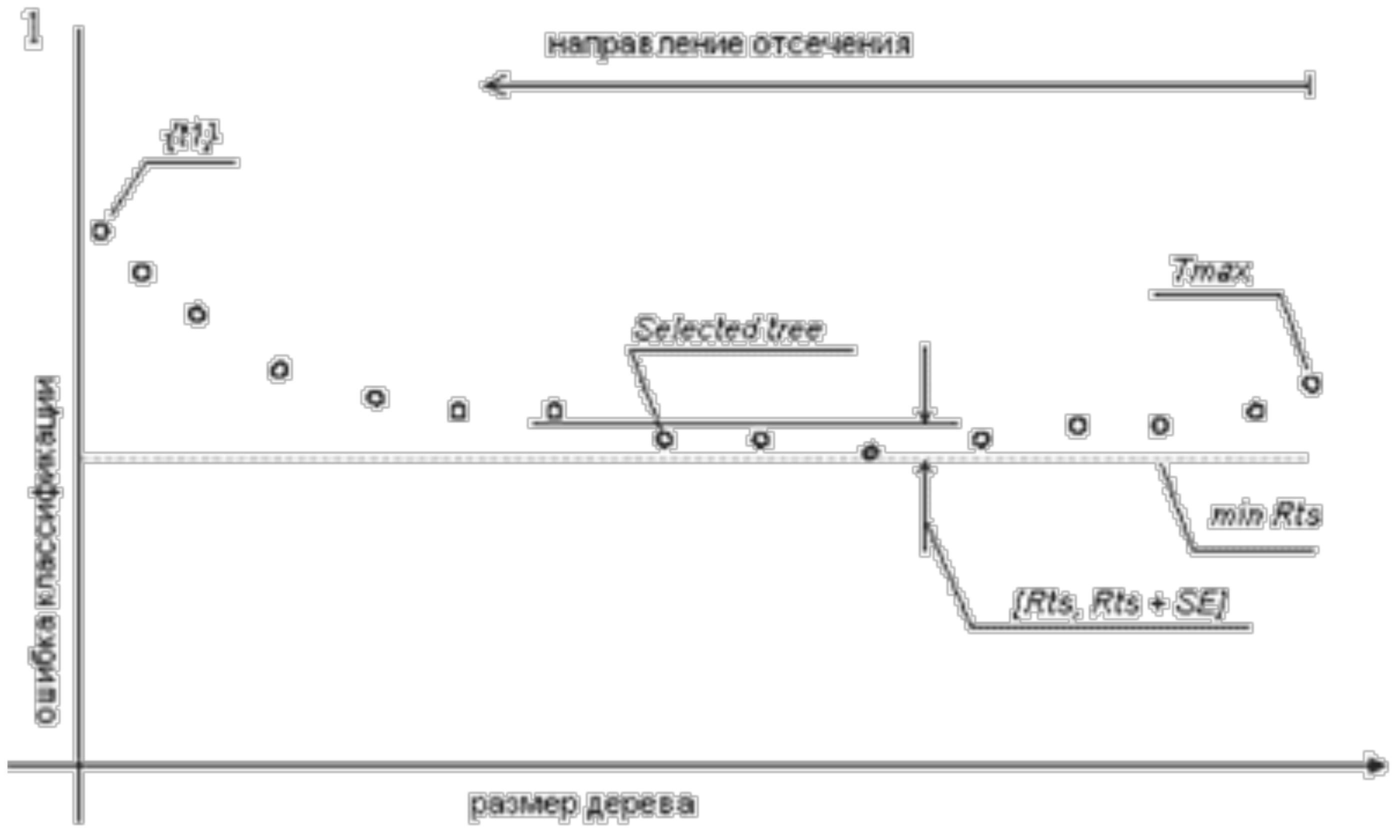
CART: V-fold cross-validation

- Как можно выбирать размер дерева:



- Как лучше выбирать размер дерева:

CART: 1-SE rule



Ссылки

- Подробней про CART:

  http://www-rohan.sdsu.edu/~jjfan/sta702/ctree.pdf

  http://scg.sdsu.edu/trees/

- XGBoost:

  https://arxiv.org/abs/1603.02754

Резюме

I. Ансамбли решающих деревьев
   a) Анализ RF и GBDT
   b) XGBoost

II. Решающие деревья
   a) Критерии информативности
   b) Прunнинг
   c) Категориальные признаки
   d) Пропущенные значения
   e) ID3, C4.5, CART

Отзывы

Отзывы о прошедших лекциях и семинарах можно и нужно оставлять здесь:

https://ml-mipt.github.io/2017part1/