Кружок по искусственному интеллекту

Семинар 2

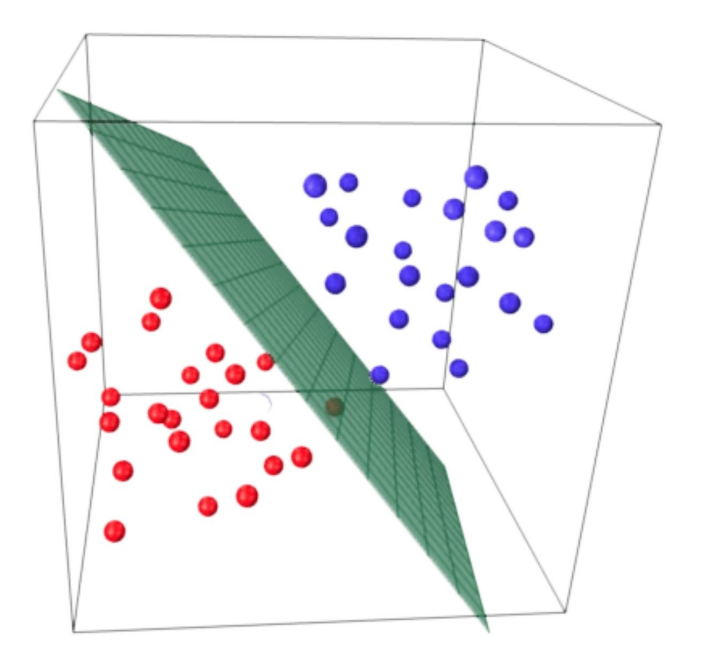
Организатор: Зубрихина Мария

Обучение с учителем

- Логистическая регрессия
- Обучение но основе решающих деревьев
- Random Forest
- К –ближайших соседей

Логистическая регрессия

- Логистическая регрессия –это линейная модель бинарной классификации.



$$\frac{p}{1-p},$$

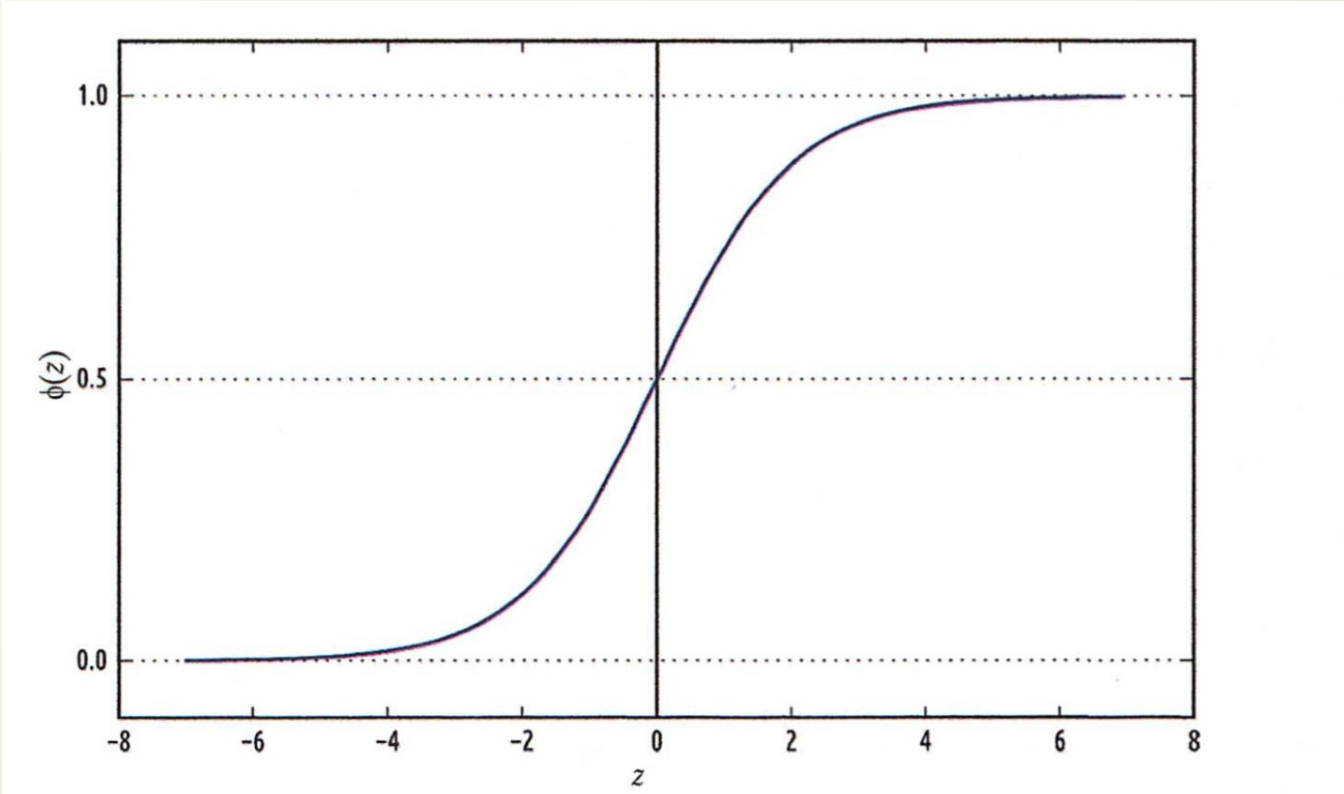
$$\text{logit}(p) = \log \frac{p}{(1-p)}.$$

$$logit\big(p(y=1\,|\,x)\big) = w_0 x_0 + w_1 x_1 + \ldots + w_m x_m = \sum_{i=0}^{m} w_i x_i = \boldsymbol{w}^T \boldsymbol{x}.$$

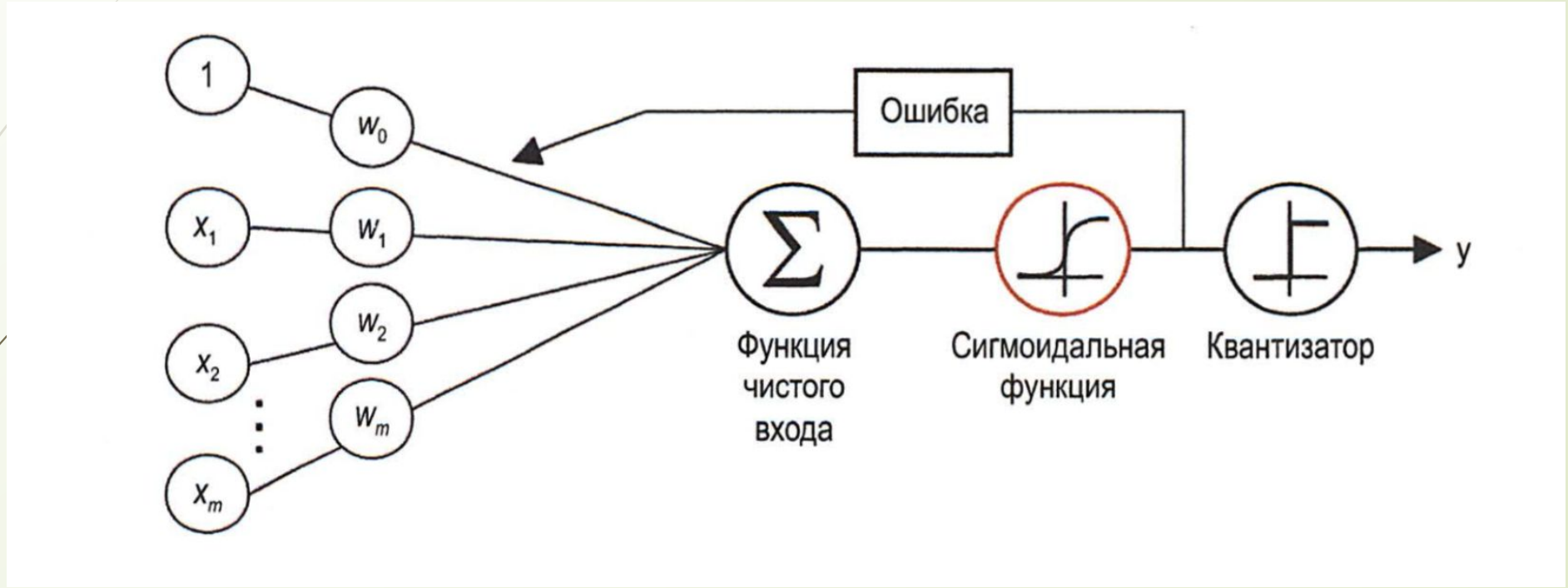
$$\phi(z) = \frac{1}{1+e^{-z}}.$$

Логистическая функция (сигмоида)

$$\phi(z) = \frac{1}{1+e^{-z}}.$$



Логистическая регрессия



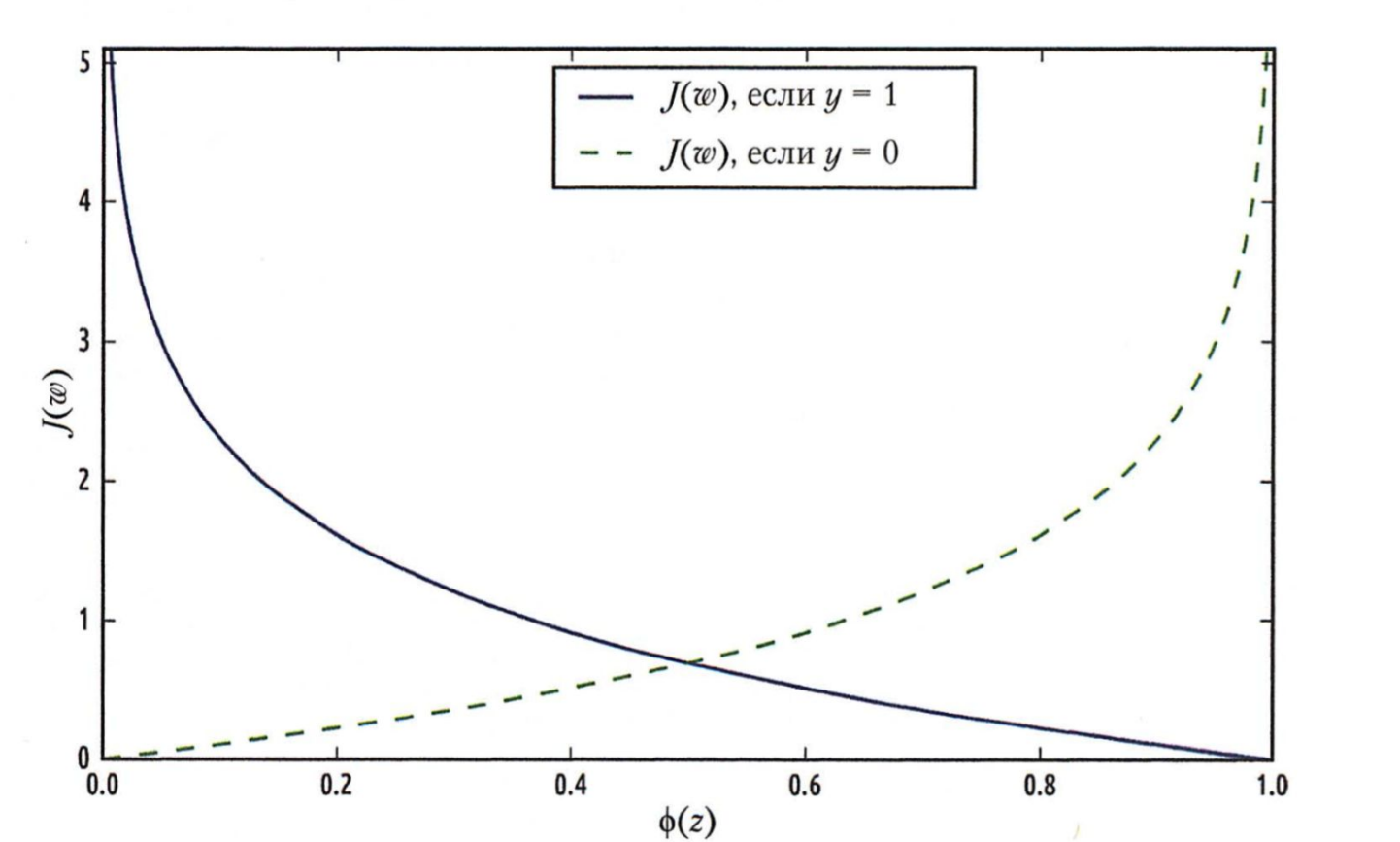
$$L(\boldsymbol{w}) = P(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{w}) = \prod_{i=1}^{n} P\left(y^{(i)} \mid x^{(i)}; \boldsymbol{w}\right) = \prod_{i=1}^{n} \left(\phi(z^{(i)})\right)^{y^{(i)}} \left(1 - \phi(z^{(i)})\right)^{1-y^{(i)}}.$$

$$w_j := w_j + \eta \sum_{i=1}^{n} \left(y^{(i)} - \phi(z^{(i)})\right) x_j^{(i)}.$$

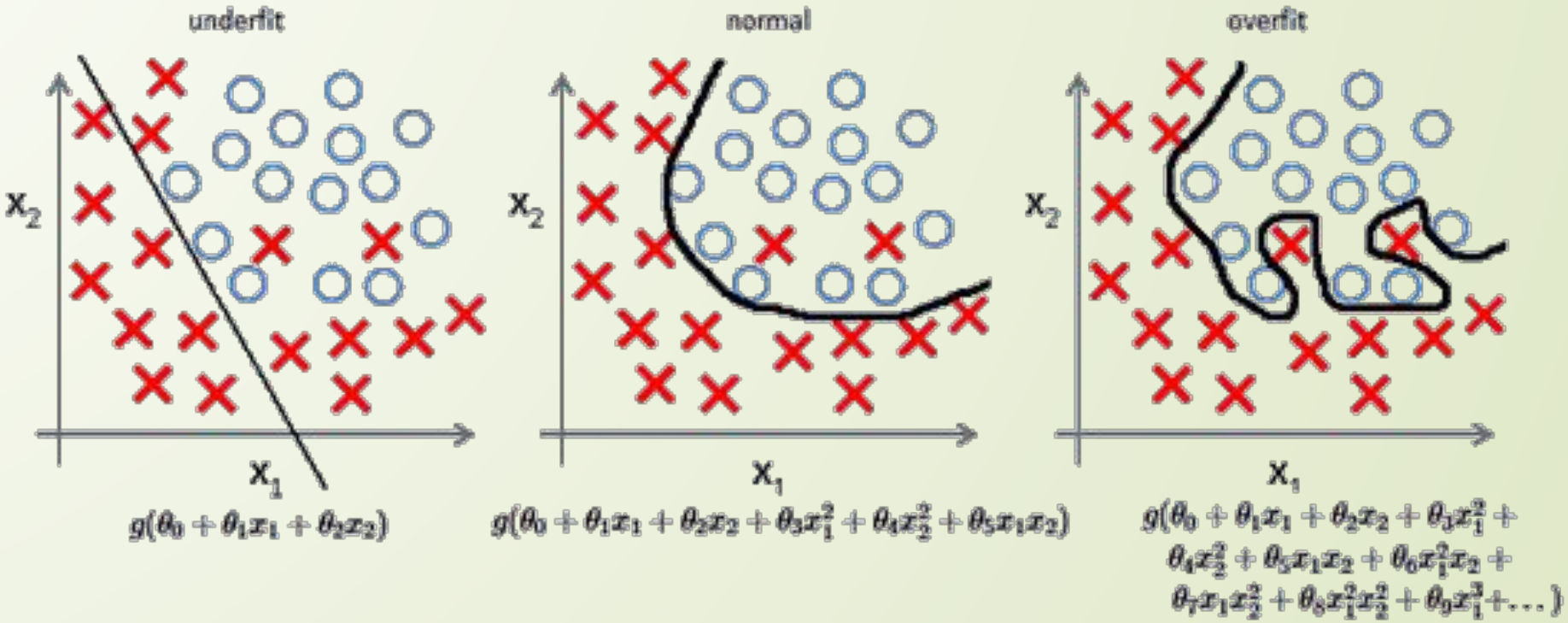
$$J(\boldsymbol{w}) = \sum_{i=1}^{n} \left[-y^{(i)} \log\left(\phi(z^{(i)})\right) - \left(1 - y^{(i)}\right) \log\left(1 - \phi(z^{(i)})\right)\right].$$

Логистическая регрессия

- функции стоимости



Переобучение



Переобучение

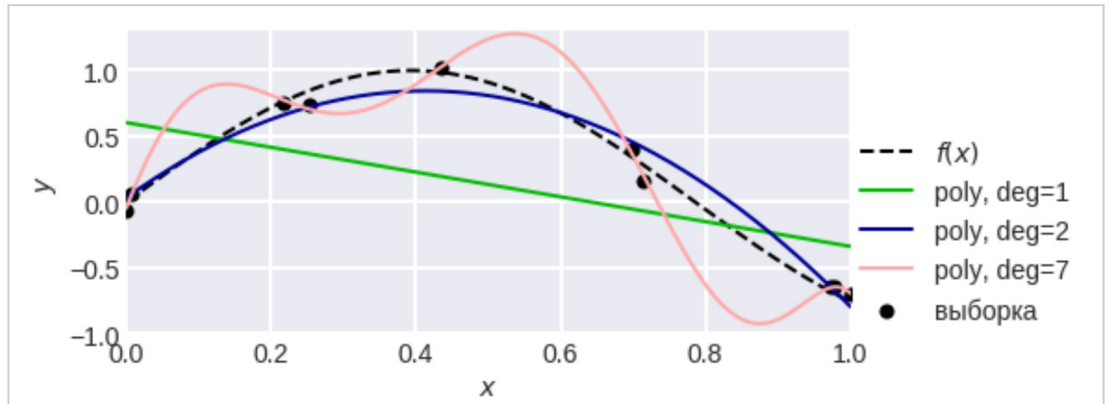


Рис. 1. Настройка полиномов различных степеней на обучающую выборку.

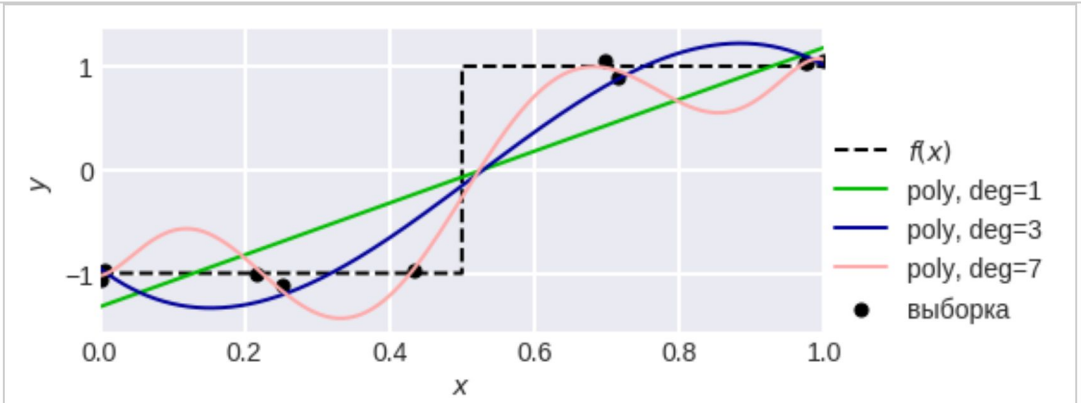


Рис. 2. Настройка полиномов различных степеней на обучающую выборку.

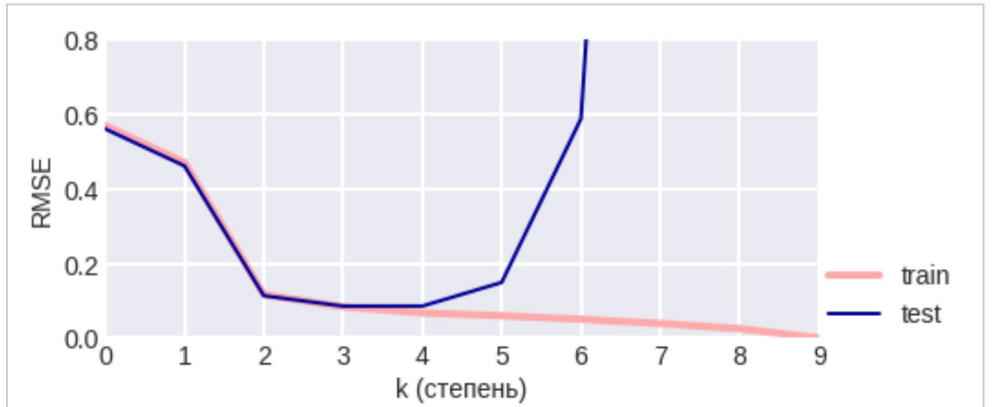


Рис. 3. Зависимость ошибки на обучении и тесте от степени полинома.

Смещение (bias) и разброс (variance)

$$y \equiv y(x) = f(x) + \varepsilon, \varepsilon \sim \text{norm}(0, \sigma^2)$$

$$\begin{aligned}
&\mathrm{E}(y-a)^2 = \mathrm{E}(y^2 + a^2 - 2ya) = \\
&= \mathrm{E}\,y^2 - (\mathrm{E}\,y)^2 + (\mathrm{E}\,y)^2 + \mathrm{E}\,a^2 - (\mathrm{E}\,a)^2 + (\mathrm{E}\,a)^2 - 2f\,\mathrm{E}\,a = \\
&= \mathrm{D}\,y + \mathrm{D}\,a + (\mathrm{E}\,y)^2 + (\mathrm{E}\,a)^2 - 2\,\mathrm{E}\,ya = \\
&= \mathrm{D}\,y + \mathrm{D}\,a + f^2 + (\mathrm{E}\,a)^2 - 2f\,\mathrm{E}\,a = \\
&= \mathrm{D}\,y + \mathrm{D}\,a + (\mathrm{E}(f-a))^2 \equiv \sigma^2 + \text{variance}(a) + \text{bias}^2(f,a)
\end{aligned}$$

Разброс (variance) - дисперсия ответов алгоритмов D***a***
Смещение (bias) – матожидание разности между истинным ответом и выданным алгоритмом: E(***f* – *a***).

Смещение (bias) и разброс (variance)

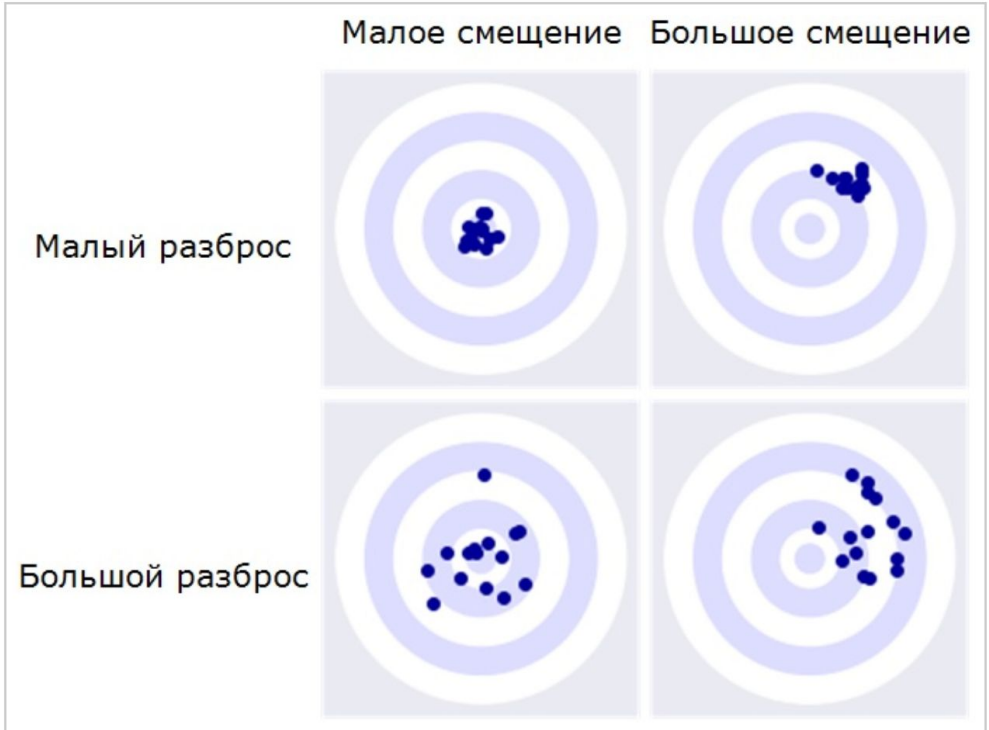


Рис. 8. Объяснение разброса и смещения на примере игры в дартс.

Переобучение

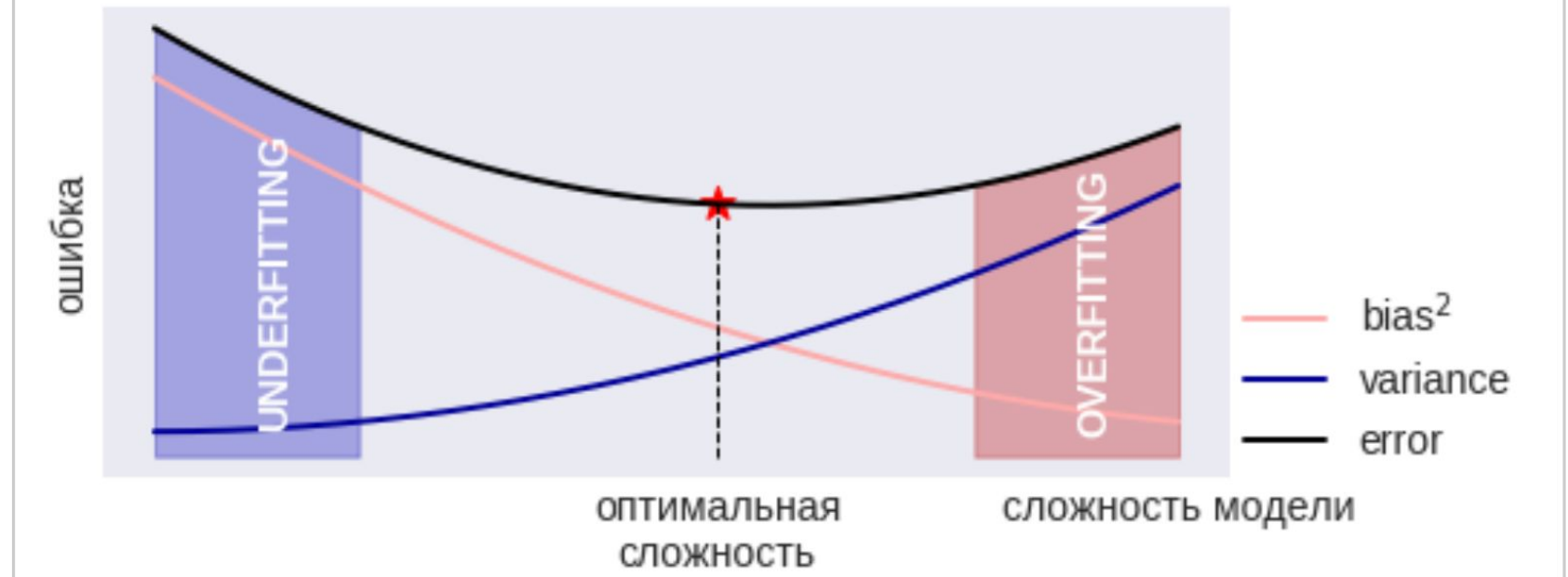


Рис. 9. Классическая иллюстрация изменения разброса и смещения.

Регуляризация

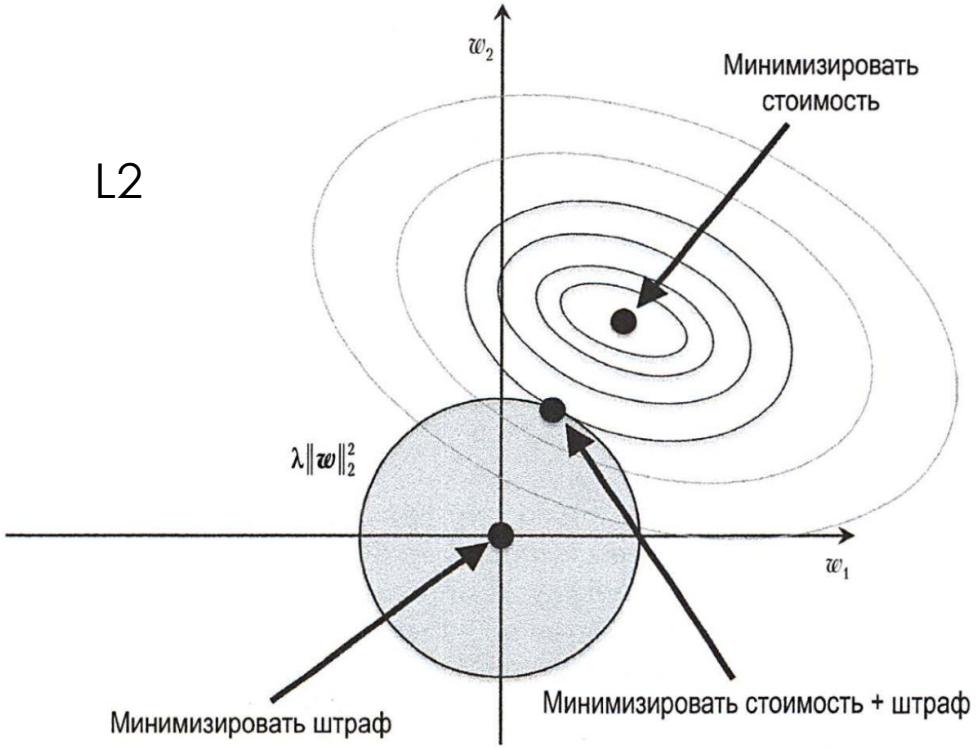
- Регуляризация - метод для обработки коллинеарности (высокой корреляции среди признаков), фильтрации шума из данных и в конечном счете предотвращения переобучения

$$L2 : \|w\|_2^2 = \sum_{j=1}^{m} w_j^2.$$

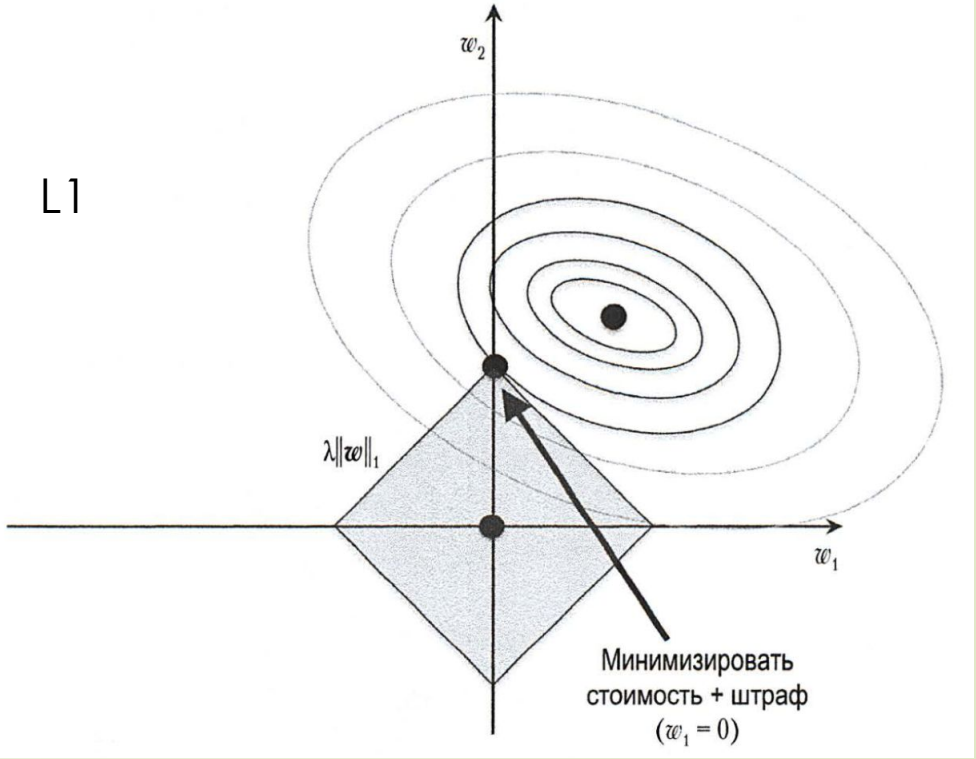
$$L1 : \|w\|_1 = \sum_{j=1}^{m} |w_j|.$$

Регуляризация

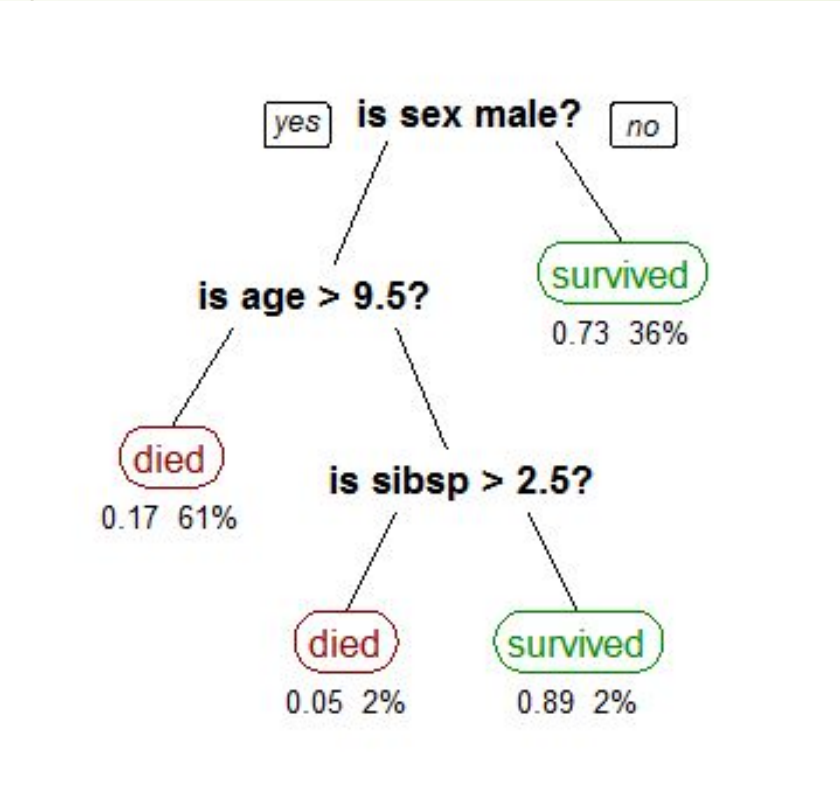
L2

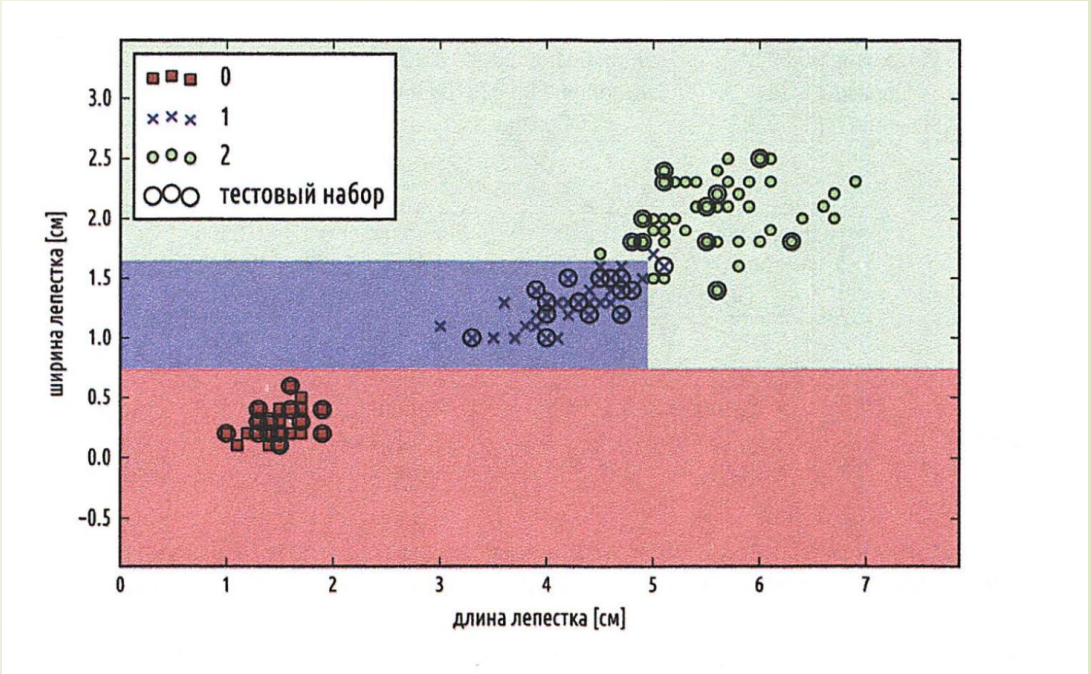


L1



Решающие деревья





Максимизация прироста информации

$$IG(D_p, f) = I(D_p) - \frac{N_{\text{левый}}}{N_p} I(D_{\text{левый}}) - \frac{N_{\text{правый}}}{N_p} I(D_{\text{правый}}).$$

D_P – это набор данных родительского узла,
$D_{\text{левый}}$ - это набор данных левого дочернего узла
$D_{\text{правый}}$ - это набор данных правого дочернего узла
N_p - общее число образцов в родительском узле
I – мера неоднородности

Меры неоднородности

- Энтропия

$$I_H(t) = -\sum_{i=1}^{c} p(i|t)\log_2 p(i|t).$$

- Мера неоднородности Джини

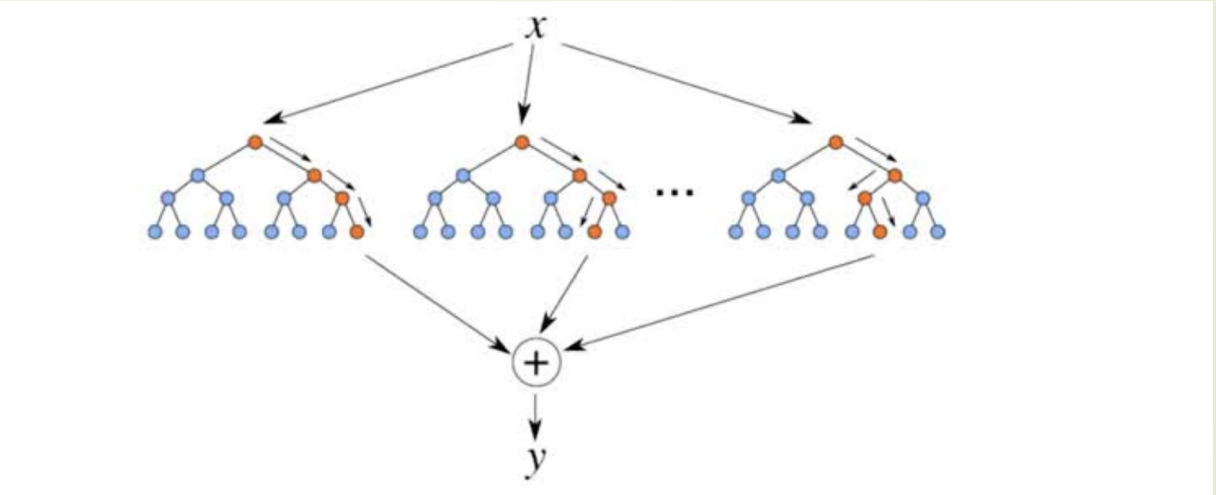
$$I_G(t) = \sum_{i=1}^{c} p(i|t)(1-p(i|t)) = 1-\sum_{i=1}^{c} p(i|t)^2.$$

- Ошибка классификации

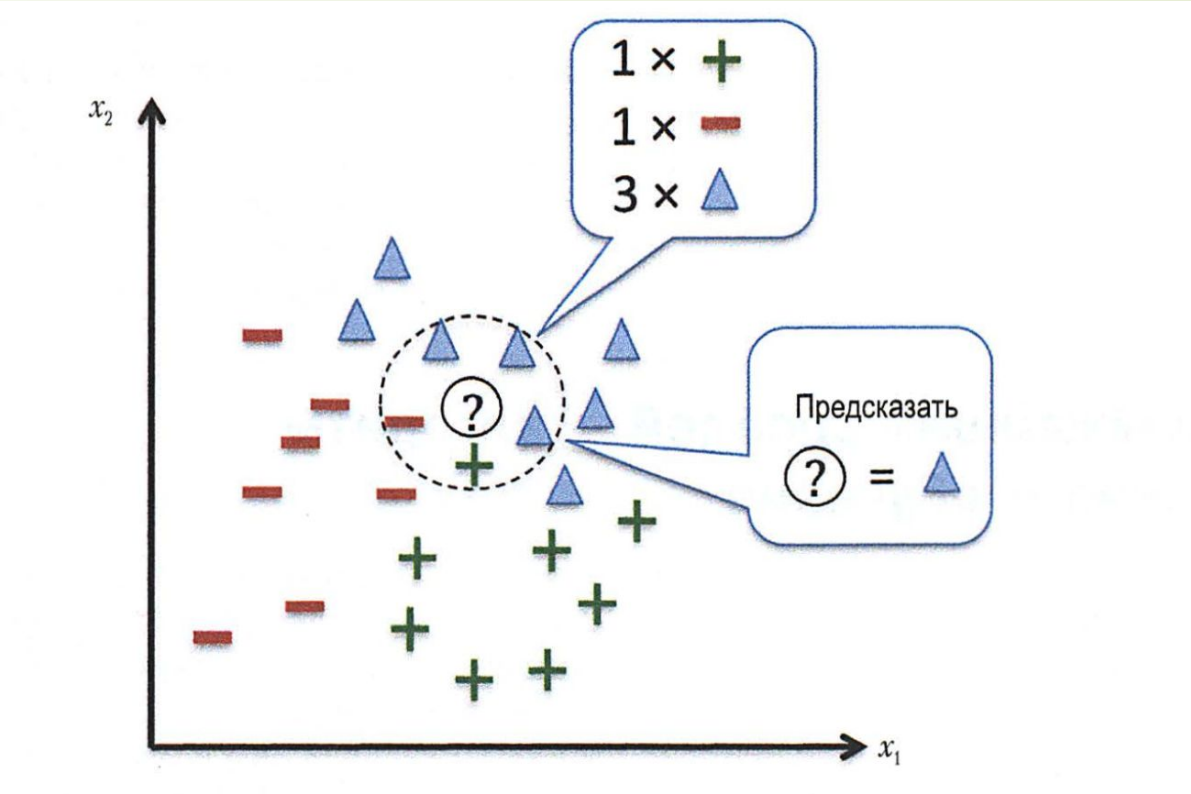
$$I_E(t) = 1 - \max\{p(i|t)\}.$$

Random Forest (Случайный лес)

- 1 шаг. Случайным образом выбрать n образцов с возвратом. (извлечь бутсрап-выборку)
- 2 шаг. Построить дерево на основе данных образцов по d признаков без возврата.
- Агрегировать прогноз на основе большинства голосов



К –ближайших соседей



Спасибо за внимание!