Logistic regression, classification

- Правдоподобие
- Метод максимального правдоподобия
- MSE
- линейная регрессия
- H0 и H1 (альтернативная)
- Sensitivity
- Specificity

Задача классификации

По признакам - X мы должны определить

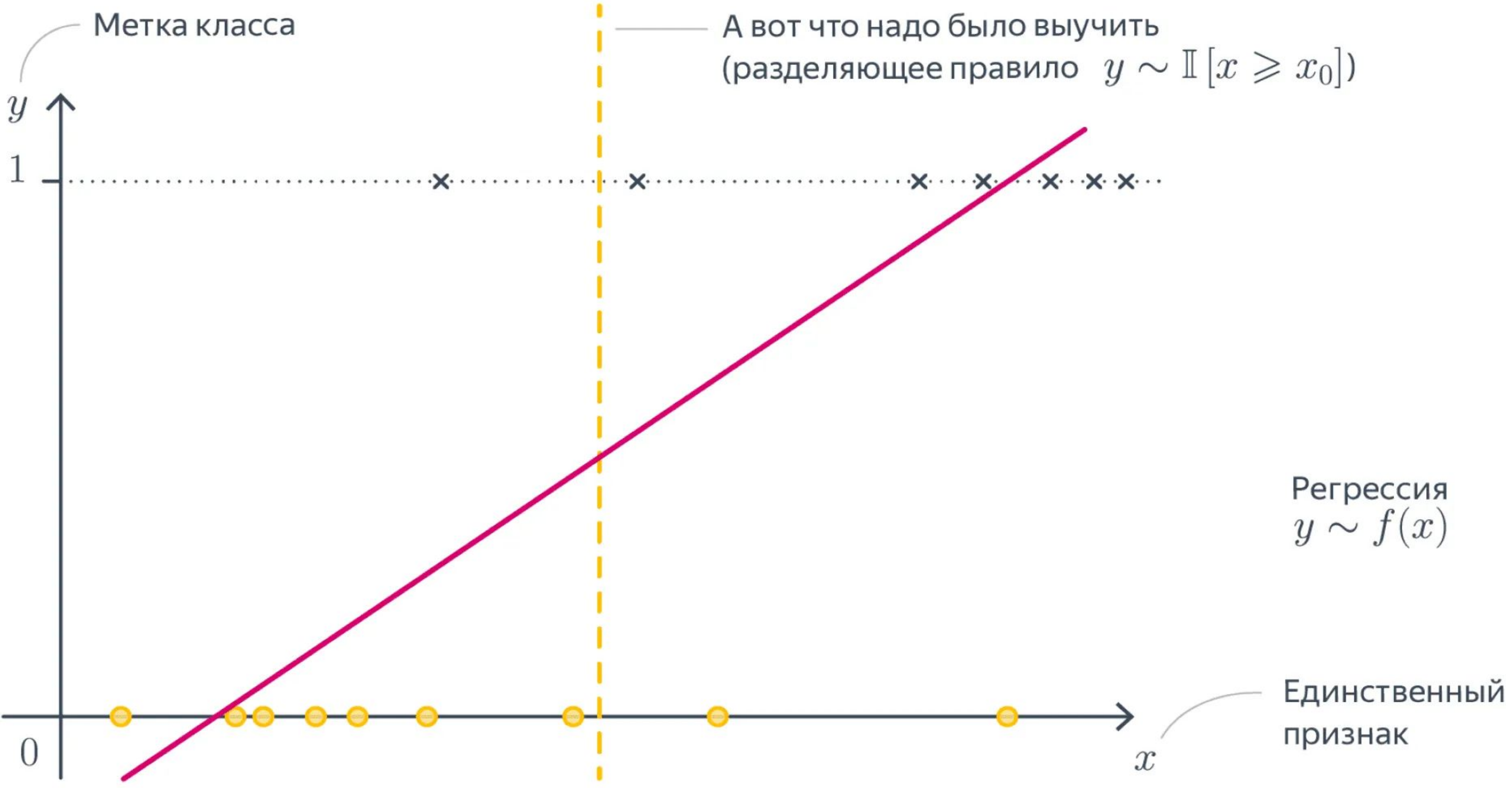
класс- C

Теперь у нас есть четкие ограничения для предсказания моделей. Они должны быть N

(0 тоже входит и иногда при бинарной классификации -1)

Как по-другому поставить задачу (более формально)?

Почему не решать её как регрессионную задачу?


Метка класса
А вот что надо было выучить
(разделяющее правило $y \sim \mathbb{I}[x \geqslant x_0]$)
y
1
0
x
Регрессия
$y \sim f(x)$
Единственный признак

Самый простой вариант линейного классификатора

!Сейчас будем говорить только о бинарной классификации!

$$c(x) = \begin{cases} 1, & \text{if } f(x) \geq 0 \\ -1, & \text{if } f(x) < 0 \end{cases}$$

Margin(отступ)

$$M_i = y_i \cdot f(x_i)$$

Главное свойство: если margin положителен, то классификация успешна

$$M_i > 0 \Leftrightarrow y_i = c(x_i)$$

$$M_i \leq 0 \Leftrightarrow y_i \neq c(x_i)$$

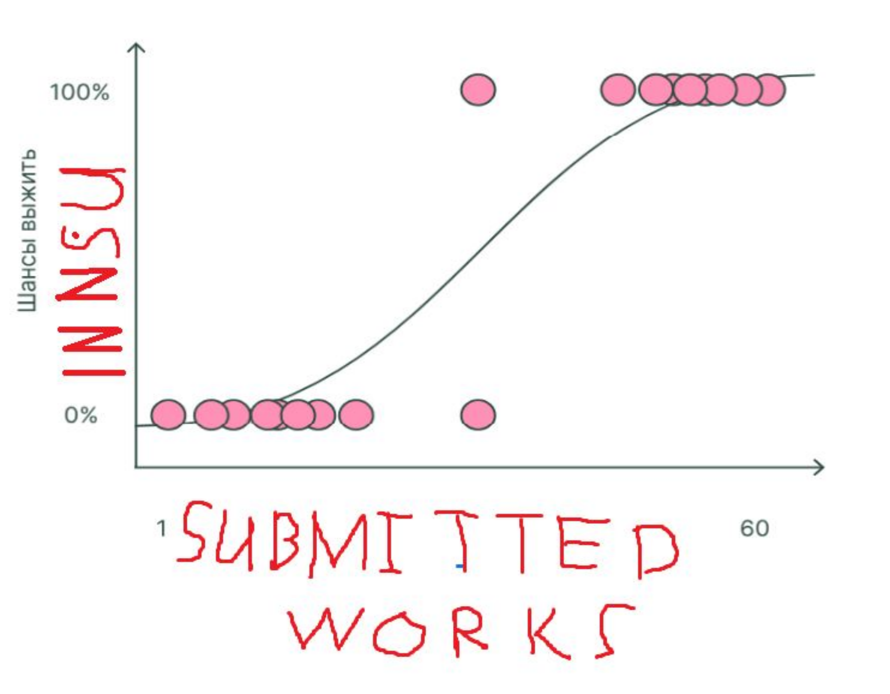
Loss функция для классификации (если нас не отпускает регрессия)

$$L_{\mathrm{MSE}} = (y_i - x_i^T w)^2 = \frac{(y_i^2 - y_i \cdot x_i^T w)^2}{y_i^2} =$$
$$= (1 - y_i \cdot x_i^T w)^2 = (1 - M_i)^2$$

Почему это плохо?

Логистическая регрессия

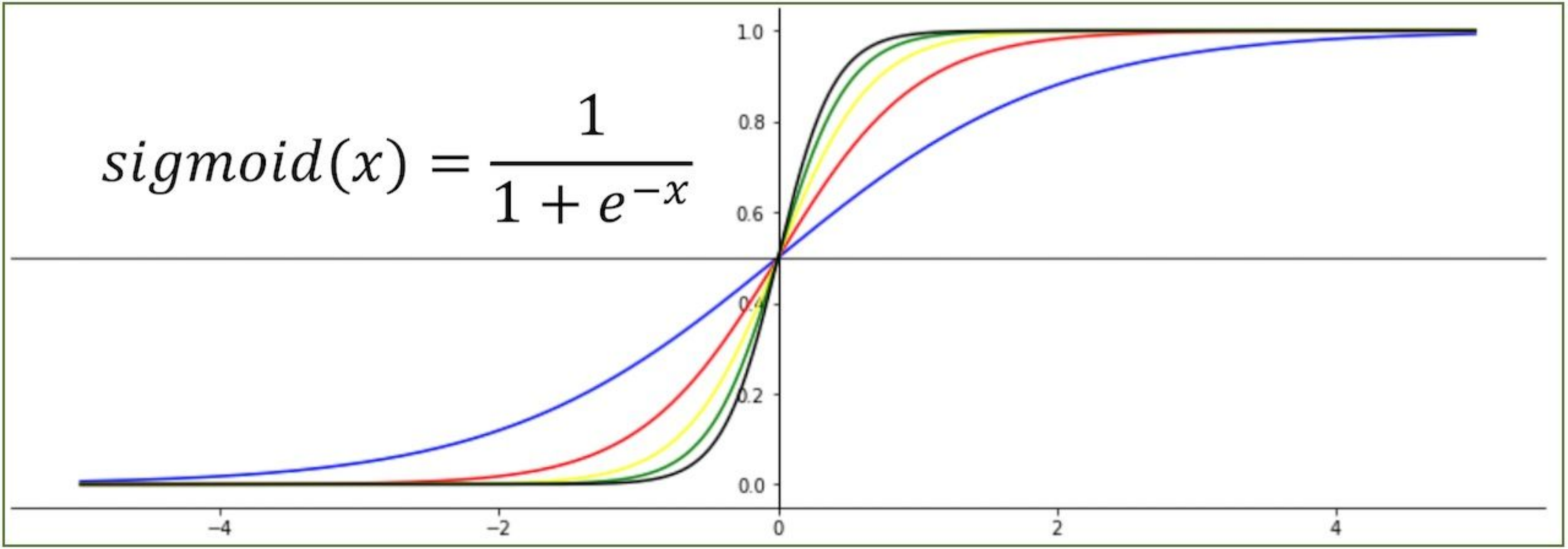
Логистическая регрессия — это тип регрессии, в которой мы оцениваем наличие связи между бинарной зависимой переменной и одной или несколькими независимыми переменными.



Сигмоида

Какая особенность у этой функции?

Почему именно она используется для классификации?



Метод максимального правдоподобия

$$\log L(w|X,Y) = \log P(X,Y|w) = \log \prod_{i=1} P(x_i, y_i|w)$$

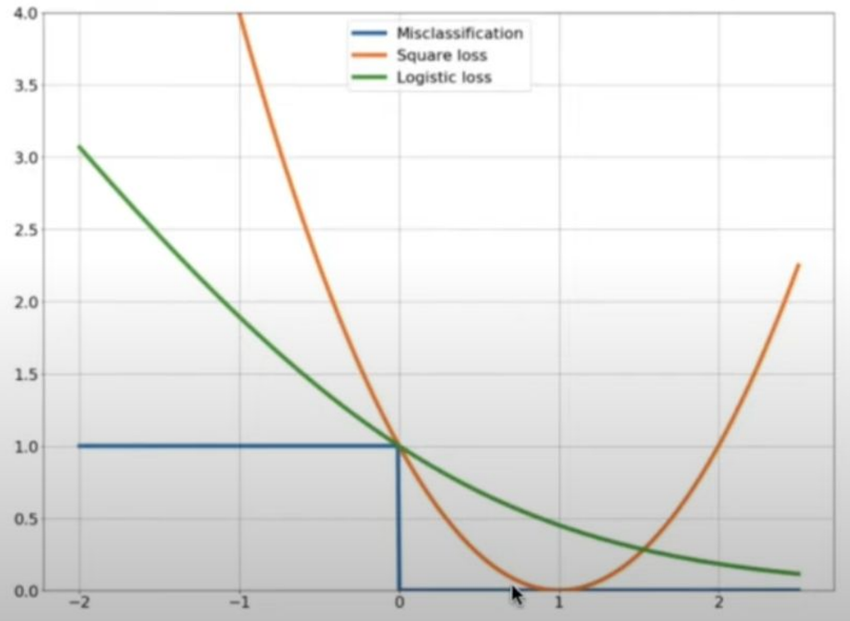
Какой шаг пропущен?

$$\log L(w|X,Y) = \sum_{i=1}^{n} \log \sigma_w(M_i) = -\sum_{i=1}^{n} \log(1 + \exp(-M_i)) \to \max_w$$

Log loss

$$L(w, X, y) = -\sum_i \left(y_i \log(\sigma(\langle w, x_i \rangle)) + (1 - y_i) \log(\sigma(-\langle w, x_i \rangle))\right)$$

$$L_{Logistic} = \log(1 + \exp(-M_i))$$



Что-то напоминает?

$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

Итак, почему же это всё-таки регрессия?

Модель на выходе нам дает следующее значение:

$$p = \sigma(\langle w, x_i \rangle)$$

Как тогда определить класс в случае бинарной классификации?

Многоклассовая классификация

В чем отличия от бинарной?

Как бороться с новыми проблемами?

One vs Rest

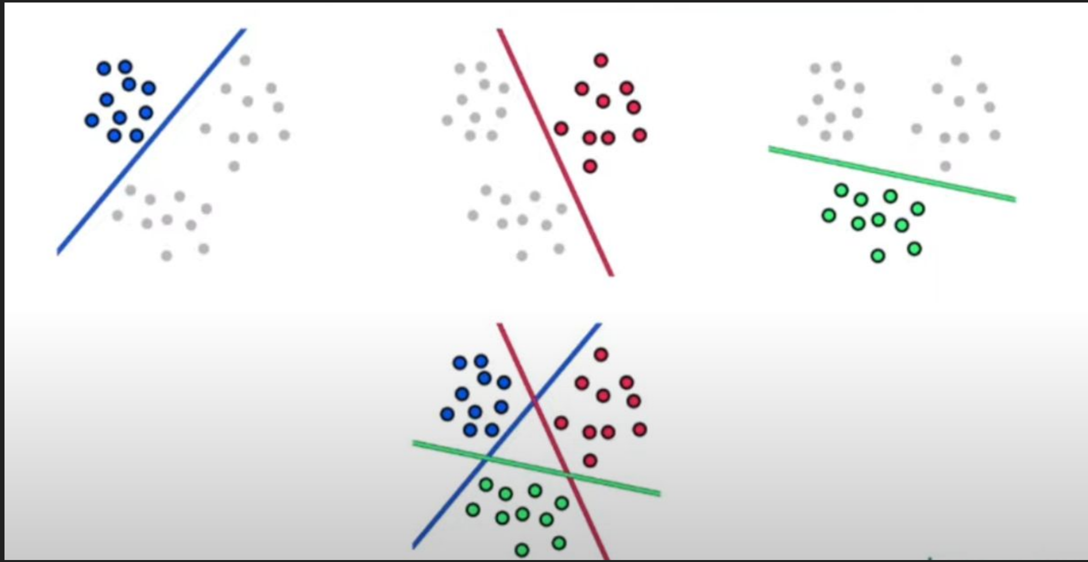
Пытаемся отличить наш класс от остальных

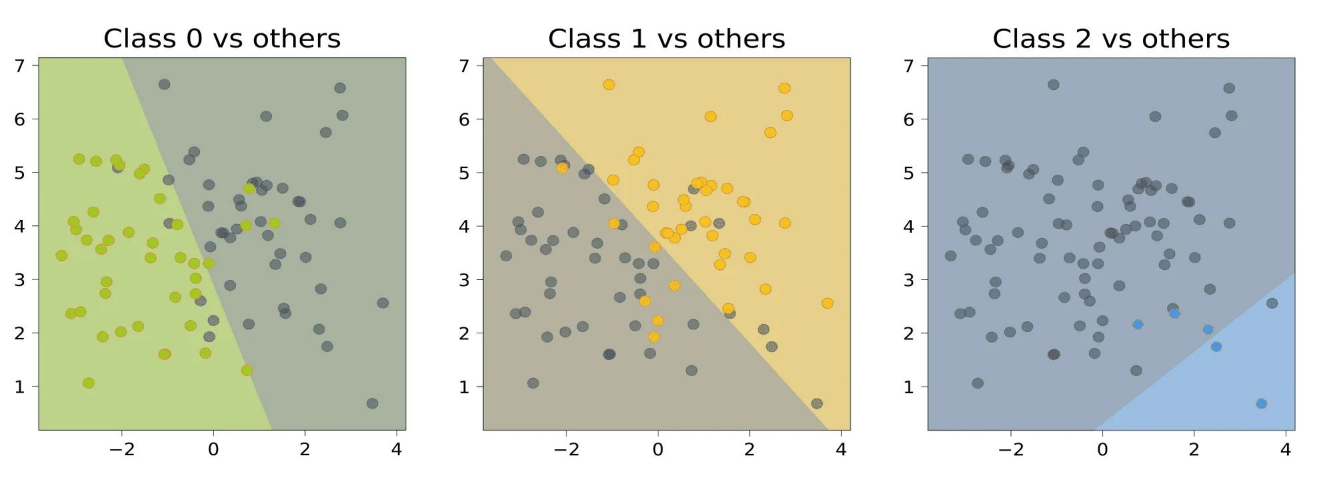
то есть > 0.5 - наш класс

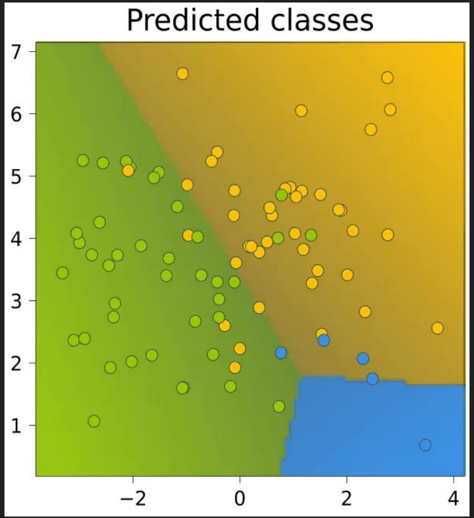
< 0.5 - другой

Какие минусы?

Что будем делать с серединой?



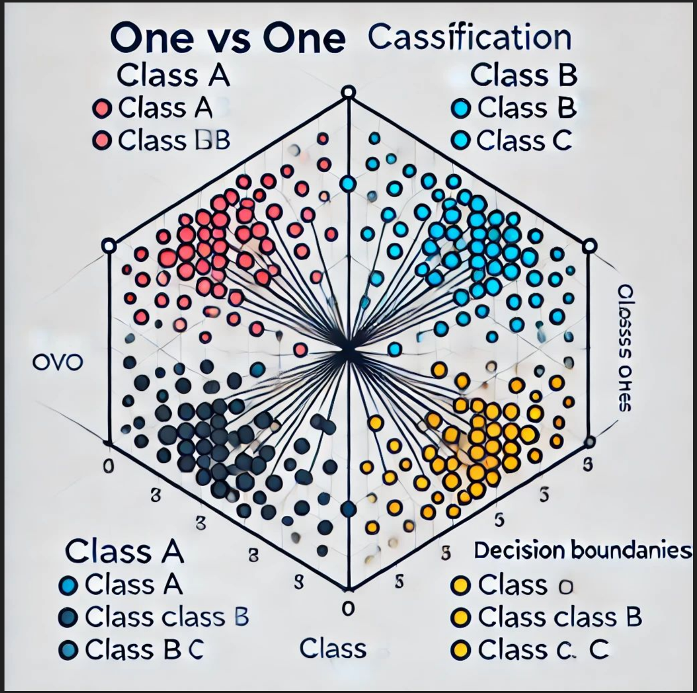

Class 0 vs others
Class 1 vs others
Class 2 vs others

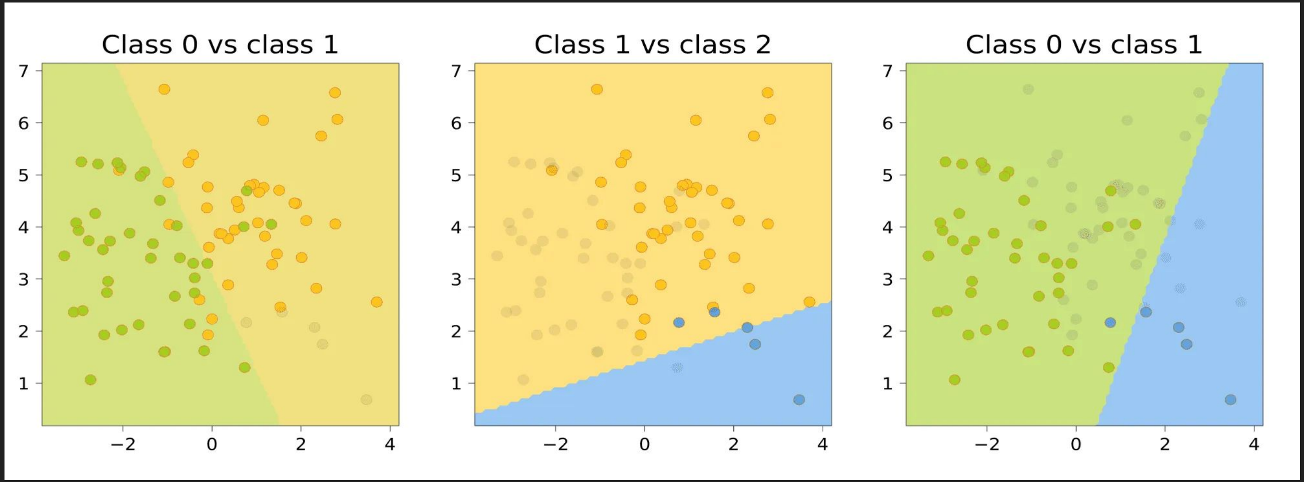

Predicted classes

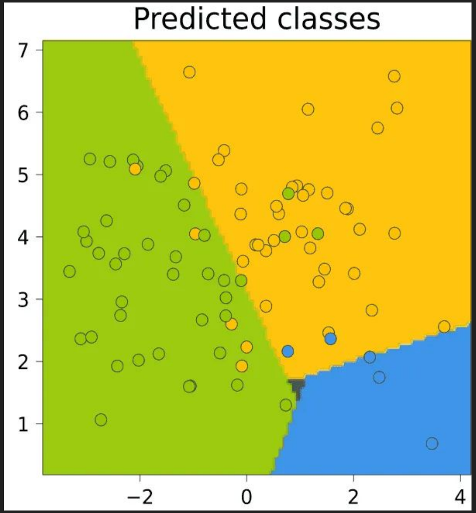
One vs One

Обучаем классификатор, чтобы он понимал где класс 0 (-1), а где 1(+1)

Делаем так для всех классов в выборке




Class 0 vs class 1
Class 1 vs class 2
Class 0 vs class 1


Predicted classes

Сравним эти 2 стратегии

One vs rest:

- Кол-во классификаторов == кол-ву классов
- Учимся на всей выборке

One vs one:

- Кол-во классификаторов == k(k-1)/2
- Учимся на подвыборках

Метрики

1. Accuracy(balanced accuracy)
2. Precision
3. Recall
4. F-score
5. ROC-curve (ROC-AUC)
6. PR (PR-AUC)
7. Confusion matrix

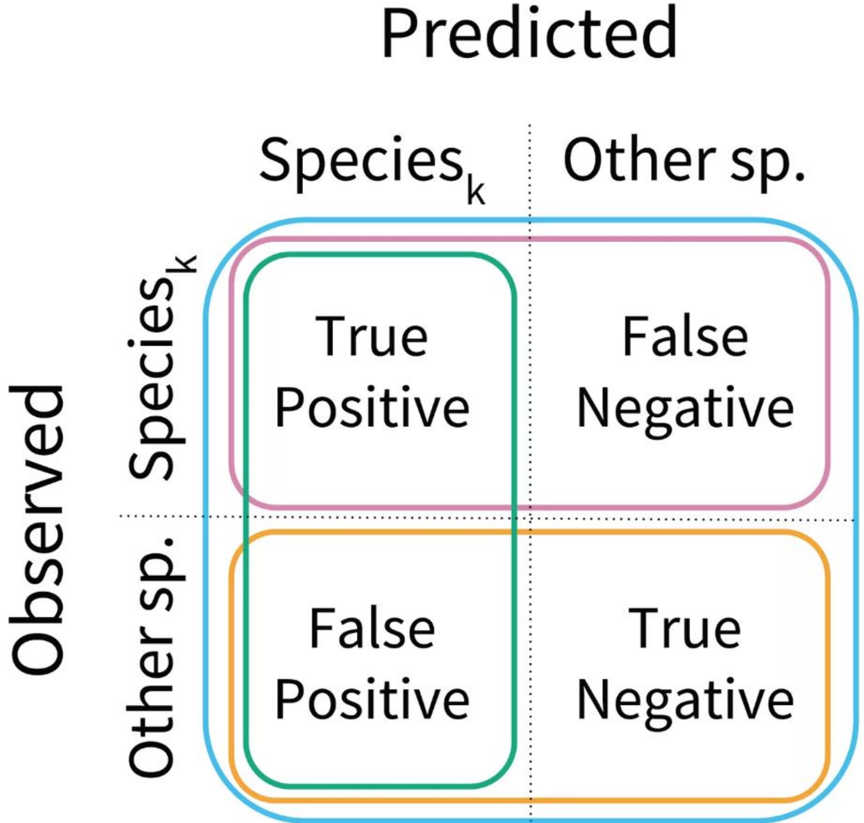
Accuracy

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}[y_i^t = y_i^p]$$

$$\text{Balanced accuracy} = \frac{1}{C}\sum_{k=1}^{C}\frac{\sum_i [y_i^t = k \text{ and } y_i^t = y_i^p]}{\sum_i [y_i = k]}$$

		Верная гипотеза	
		H_0	H_1
Результат применения критерия	H_0	H_0 верно принята	H_0 неверно принята (Ошибка *второго* рода)
	H_1	H_0 неверно отвергнута (Ошибка *первого* рода)	H_0 верно отвергнута

Precision и Recall



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

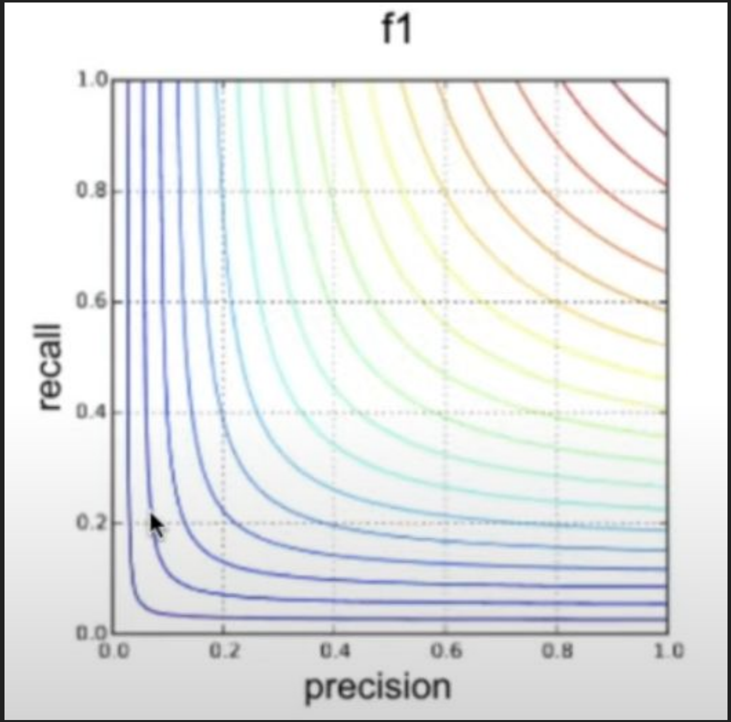
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

F-score

$$F_1 = \frac{2}{\text{precision}^{-1} + \text{recall}^{-1}} = 2\,\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

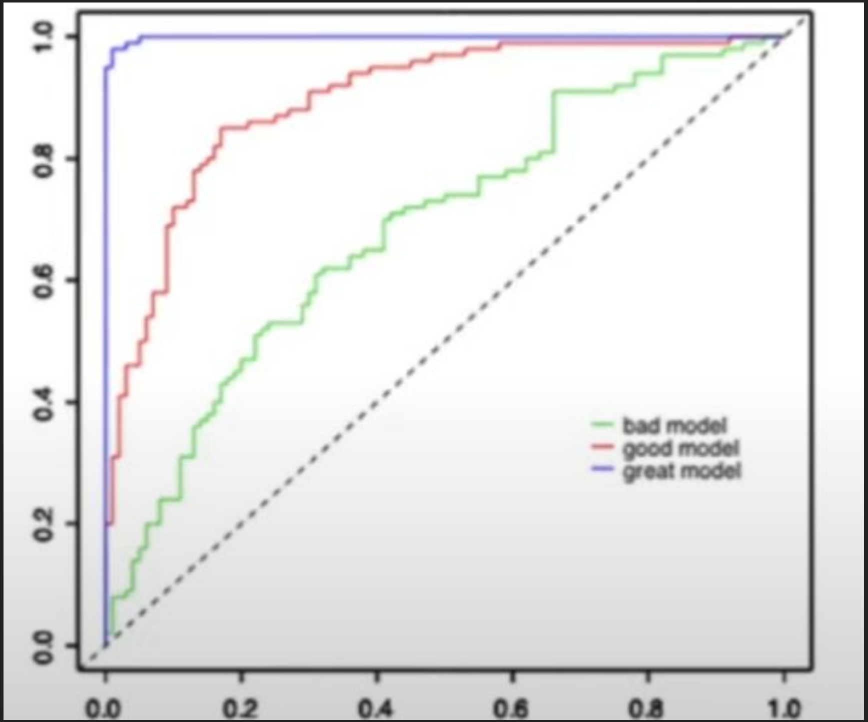
$$F_\beta = (1 + \beta^2)\frac{\text{precision} \cdot \text{recall}}{\beta^2\,\text{precision} + \text{recall}}$$



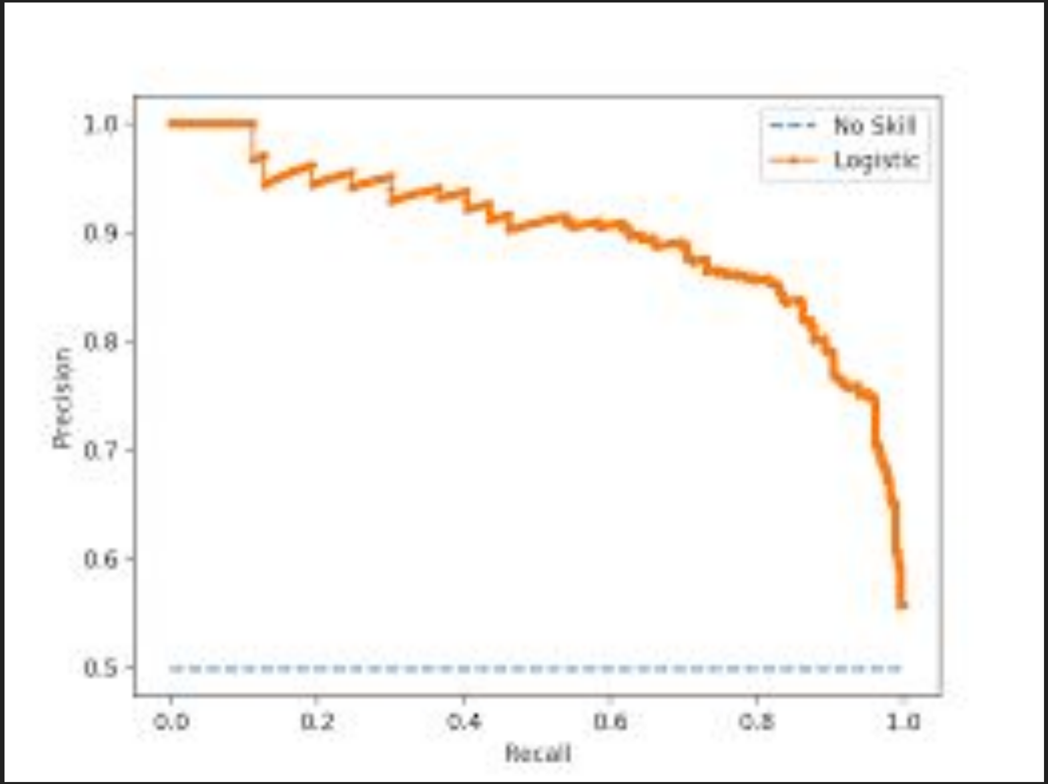
ROC - AUC

$$\mathrm{FPR} = \frac{FP}{FP + TN}$$

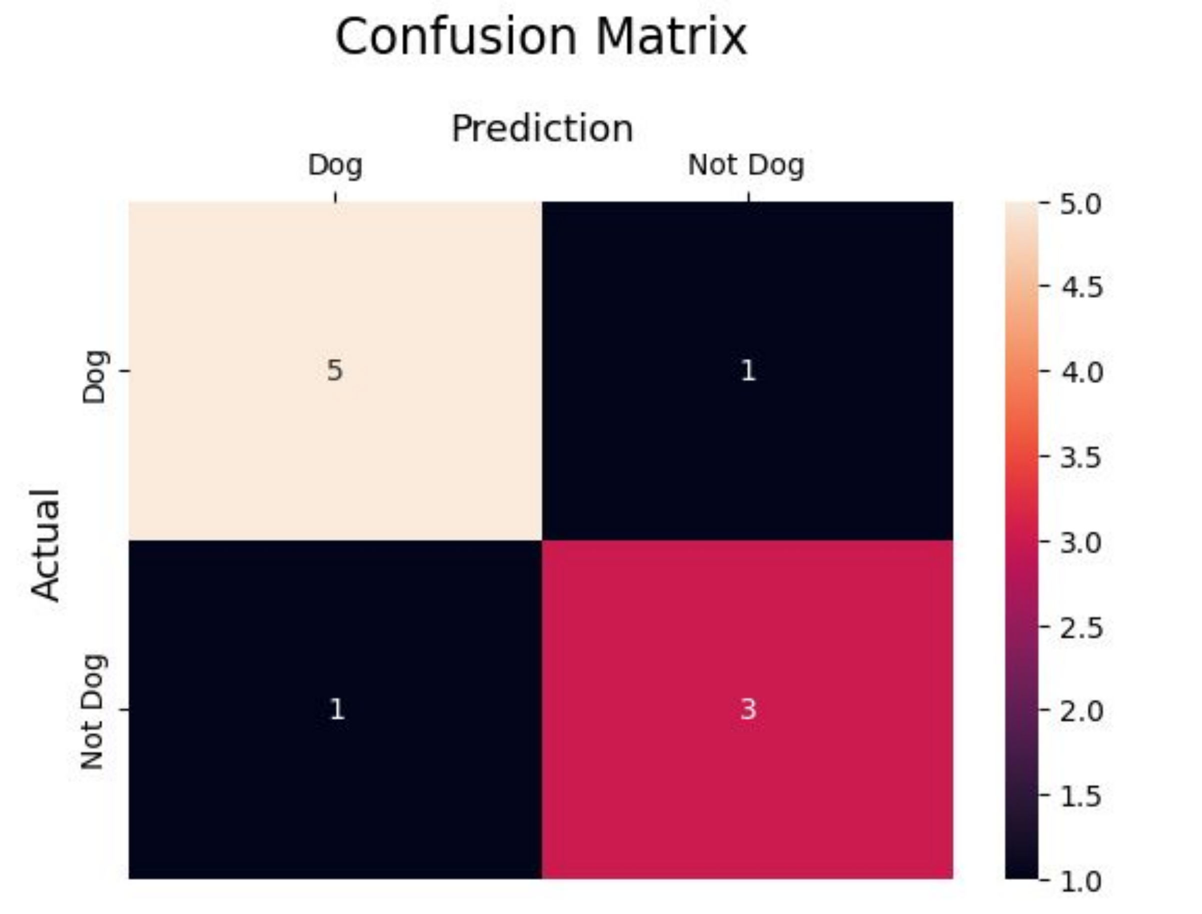
$$\mathrm{TPR} = \frac{TP}{TP + FN} (= \mathrm{Recall})$$



PR (PR-AUC)



No Skill
Logistic
1.0
0.9
0.8
0.7
0.6
0.5
Precision
0.0
0.2
0.4
0.6
0.8
1.0
Recall


Confusion Matrix
Prediction
Dog
Not Dog
Actual
Dog
Not Dog
5
1
1
3
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0