Machine Learning Course

Lecture 3: Linear classification

MIPT, 2019

Outline

1. Linear regression recap
2. Linear classification
3. Margin in linear classification
4. Loss functions
5. Gradient descent recap
6. Logistic regression
7. Quality functions in classification

Linear regression

Observed objects: $(x^i, y^i), i = 1, \dots, n$

$$x^i \in R^p, y^i \in R$$

Linear model: $f(x) = w_1 x_1 + \dots + w_p x_p = x^T w$

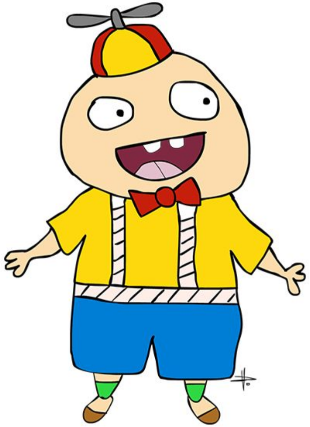
Missed free term?

$$\begin{aligned}
(x_1, \dots, x_p) &\rightarrow (1, x_1, \dots, x_p) \\
(w_1, \dots, w_p) &\rightarrow (w_0, w_1, \dots, w_p) \\
p &\rightarrow p+1 \\
n &\rightarrow n \longleftarrow \text{most important!}
\end{aligned}$$

Linear regression

$$f(x) = w_0 + w_1 x_1 + \cdots + w_p x_p$$

Schoolboy's regression



Matrix form of data:

$$X^T = [x^1, \ldots, x^n], X \in R^{n \times p}$$

$$Y^T = [y^1, \ldots, y^n], Y \in R^n$$

Real men's linear model

$$f_w(X) = Xw \approx Y$$

How to choose weights?

$$\text{Empirical risk} = \sum_{\text{by objects}} \text{Loss on object} \to \min_{\text{model params}}$$

$$Q(X) = \sum_{i=1}^{n} L(y^i, f_w(x^i)) \to \min_w$$

Loss functions

$$\text{MSE:} \quad L(y_t, y_p) = (y_t - y_p)^2$$

$$\text{MAE:} \quad L(y_t, y_p) = |y_t - y_p|$$

For MSE closed form solution exists

$$Q_{\mathrm{MSE}}(X) = \sum_{i=1}^{n} (y^i - f_w(x^i))^2 = ||Y - Xw||^2 \to \min_w$$

$$w^* = (X^T X)^{-1} X^T Y$$

Gauss-Markov theorem:

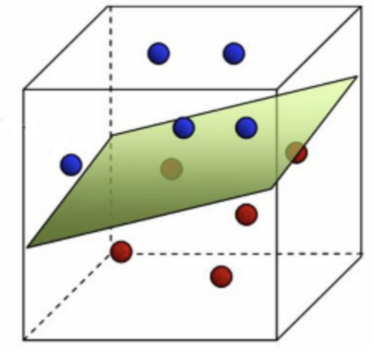
Minimizing MSE loss gives
Best Linear Unbiased Estimation (BLUE)

Linear classification

$$a(x) = \begin{cases} 1, \text{ if } f(x) > 0 \\ -1, \text{ if } f(x) < 0 \end{cases}$$

$$f(x) = \langle w, x \rangle$$

Geometrical interpretation:
Linearly separable case



Margin

Denote algorithm $a(x) = sign\{f(x)\}$

Let's call $M_i = y_i f(x_i)$ algorithm *margin* on object x_i .

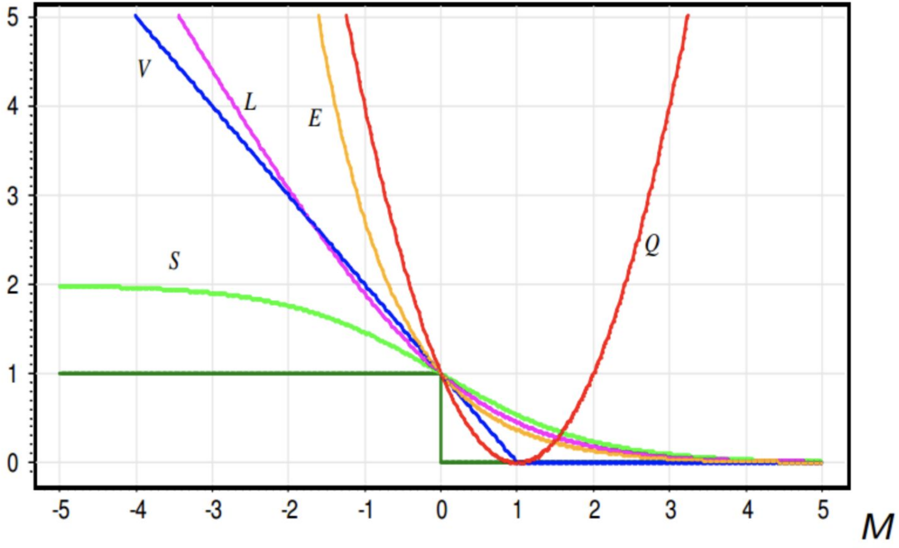
$$M_i \leq 0 \Leftrightarrow y_i \neq a(x_i)$$
$$M_i > 0 \Leftrightarrow y_i = a(x_i)$$

Loss functions

$$Q(w) = \sum_{i=1}^{\ell} [M_i(w) < 0] \leqslant \widetilde{Q}(w) = \sum_{i=1}^{\ell} \mathscr{L}(M_i(w)) \to \min_w;$$

Smoothed empirical risk

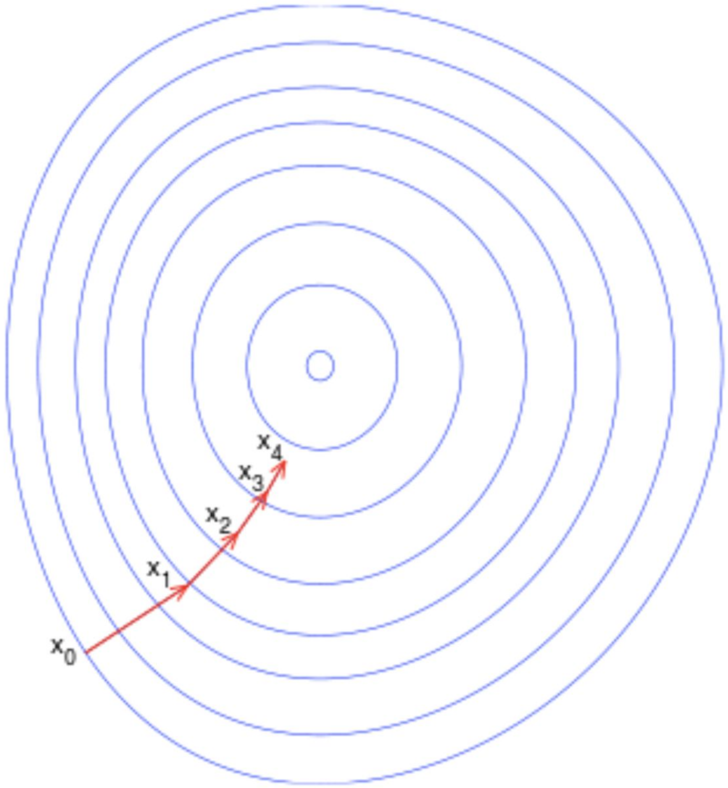
Loss function



$$Q(M) = (1 - M)^2$$
$$V(M) = (1 - M)_+$$
$$S(M) = 2(1 + e^M)^{-1}$$
$$L(M) = \log_2(1 + e^{-M})$$
$$E(M) = e^{-M}$$

Loss functions

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma_n \nabla F(\mathbf{x}_n),\ n \geq 0.$$



$$\nabla_w \tilde{Q} = \sum_{i=1}^{l} \nabla L(M_i)$$

$$\nabla \tilde{Q} = \sum_{i=1}^{l} L'(M_i) \frac{\partial M_i}{\partial w}$$

$$\frac{\partial M_i}{\partial w} = y_i x_i$$

$$\nabla \tilde{Q} = \sum_{i=1}^{l} y_i x_i L'(M_i)$$

$$\boxed{w_{n+1} = w_n - \gamma_n \sum_{i=1}^{l} y_i x_i L'(M_i)}$$

Logistic regression

$y_i \in \{0, 1\}$

$$Q = -\sum_{i=1}^{\ell} y_i \ln p_i + (1 - y_i) \ln(1 - p_i) \to \min_w$$

$$p_i = \sigma(\langle w, x_i \rangle) = \frac{1}{1 + e^{-\langle w, x_i \rangle}} = P(y = 1|x)$$

logistic loss

L1 or L2 regularization terms are usually used along the *logistic loss* function.
The optimization problem is solved by SGD or Newton-Raphson's method.

Logistic regression optimization problem

$$Q = -\sum_{i=1}^{\ell} y_i \ln \frac{1}{1+e^{-\langle w, x_i \rangle}} + (1-y_i) \ln \frac{1}{1+e^{\langle w, x_i \rangle}} \to \min_w$$

$$-y_i \ln \frac{1}{1+e^{-\langle w, x_i \rangle}} - (1-y_i) \ln \frac{1}{1+e^{\langle w, x_i \rangle}} = \begin{cases} \ln\left(1+e^{-\langle w, x_i \rangle}\right), y_i = 1 \\ \ln\left(1+e^{\langle w, x_i \rangle}\right), y_i = 0 \end{cases}$$

$$Q = \sum_{i=1}^{\ell} \underbrace{\ln\left(1+e^{-y_i \langle w, x_i \rangle}\right)}_{L(M) = \ln(1+e^{-M_i})} \to \min_w \qquad y_i \in \{-1, 1\}$$

Quality functions in classification

- Accuracy
- Precision
- Recall
- F-score
- ROC-curve, ROC-AUC
- PR-curve

Accuracy

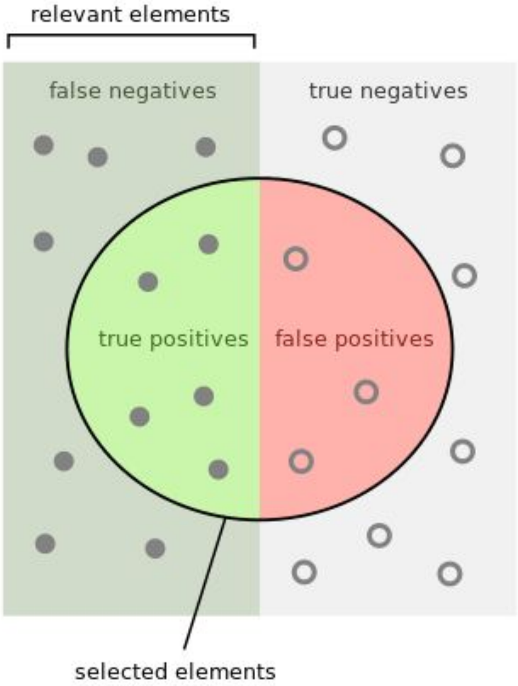
Number of right classifications

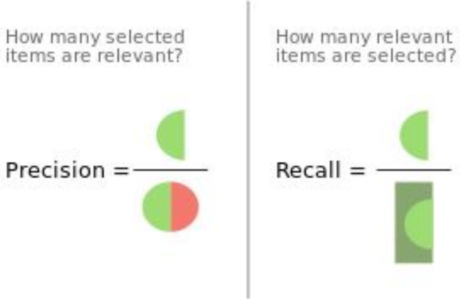
target: 1 0 1 0 0 0 0 1 0 0

predicted: 0 0 1 0 0 0 0 1 1 0

accuracy = 8/10 = 0.8

Precision and recall





		Actual Class	
		Yes	No
Predicted Class	Yes	**T**rue **P**ositive	**F**alse **P**ositive
	No	**F**alse **N**egative	**T**rue **N**egative

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

F-score

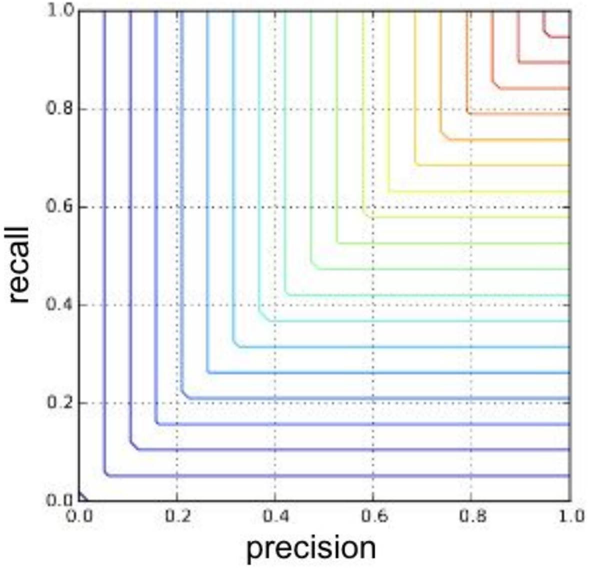
Harmonic mean of precision and recall.
Closer to the smallest one.

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

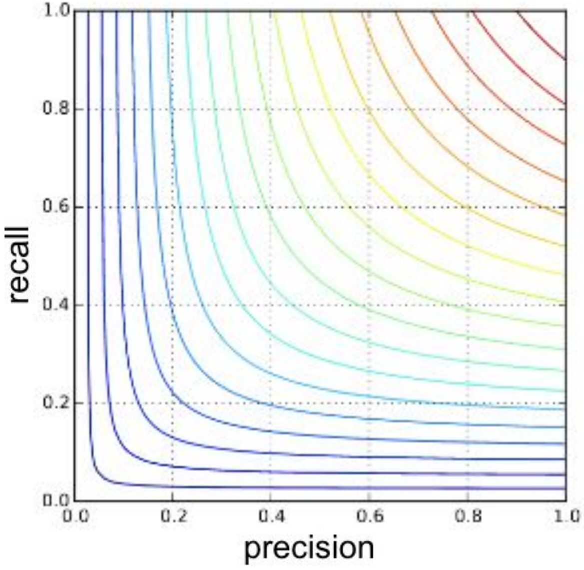
$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}$$

F-score

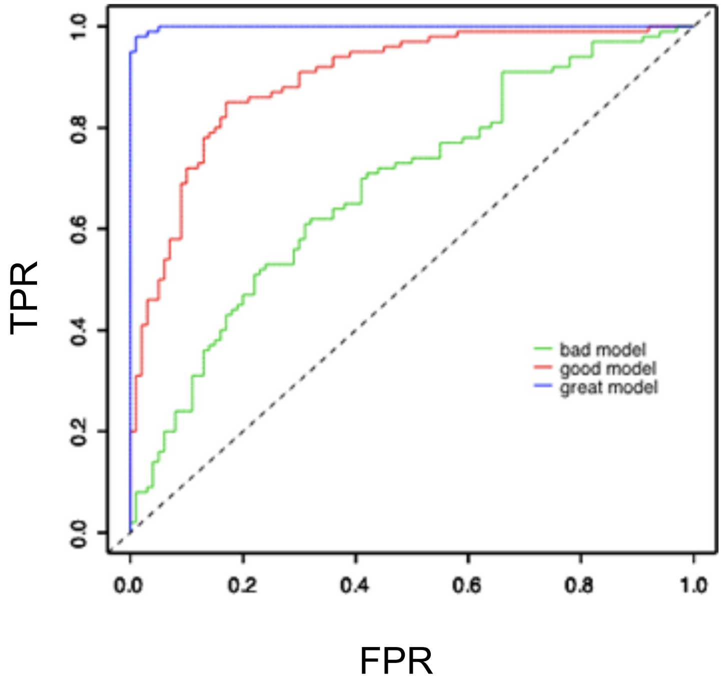
min(precision, recall)



f1



ROC - receiver operating characteristic

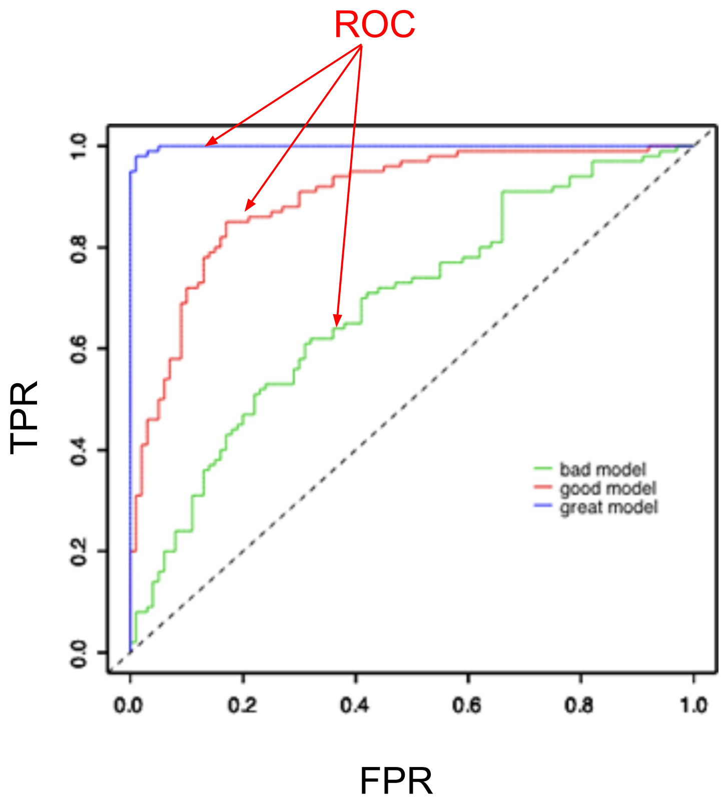


		Actual Class	
		Yes	No
Predicted Class	Yes	**T**rue **P**ositive	**F**alse **P**ositive
	No	**F**alse **N**egative	**T**rue **N**egative

$$TPR = \frac{True\ positives}{True\ positives + False\ negatives}$$

$$FPR = \frac{False\ positives}{False\ positives + True\ negatives}.$$

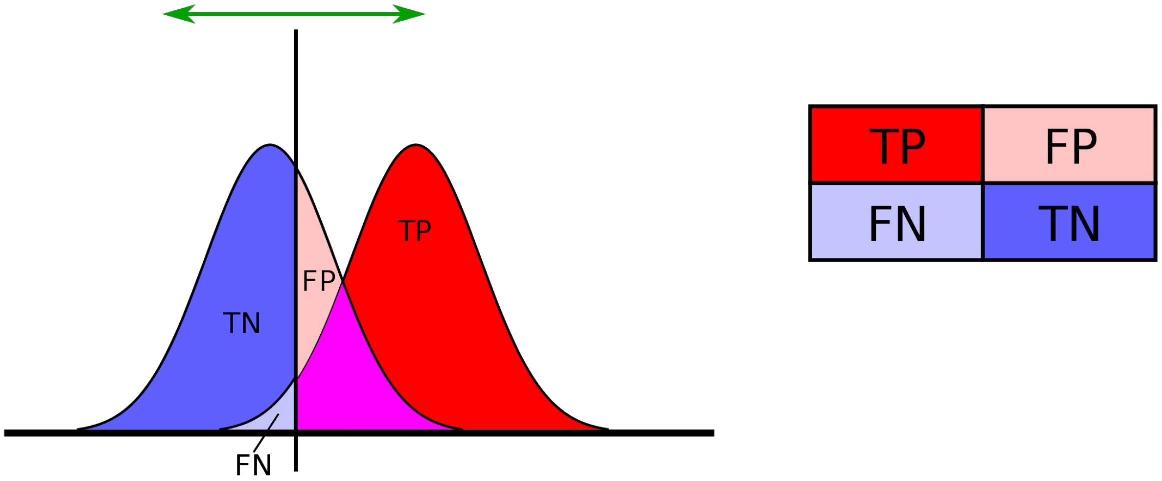
ROC

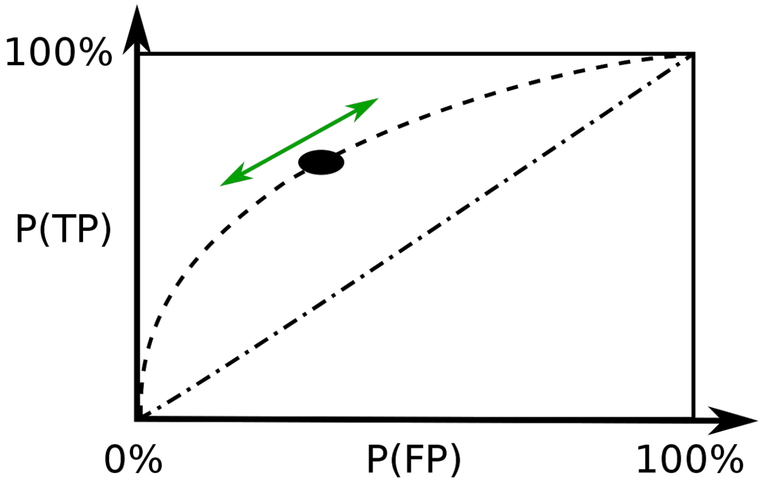


		Actual Class	
		Yes	No
Predicted Class	Yes	**T**rue **P**ositive	**F**alse **P**ositive
	No	**F**alse **N**egative	**T**rue **N**egative

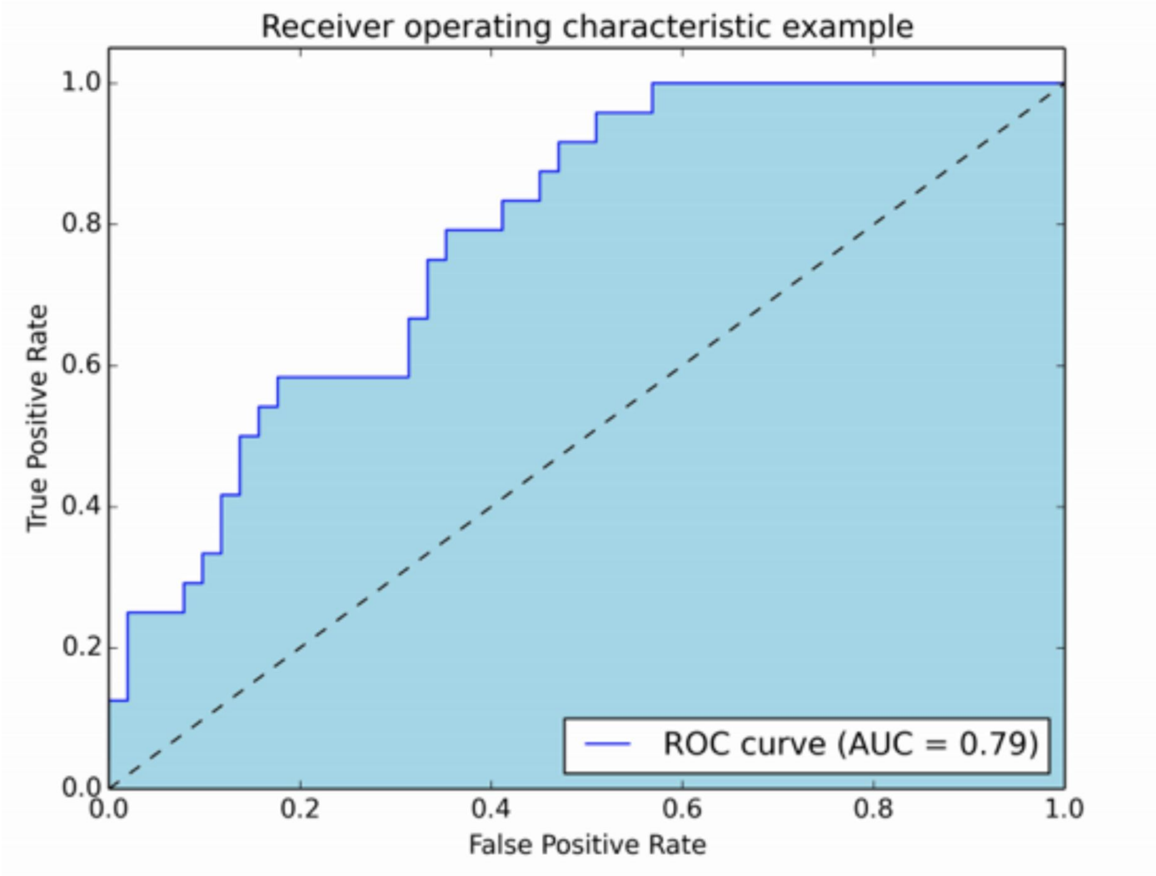
$$TPR = \frac{True\ positives}{True\ positives + False\ negatives}$$

$$FPR = \frac{False\ positives}{False\ positives + True\ negatives}.$$

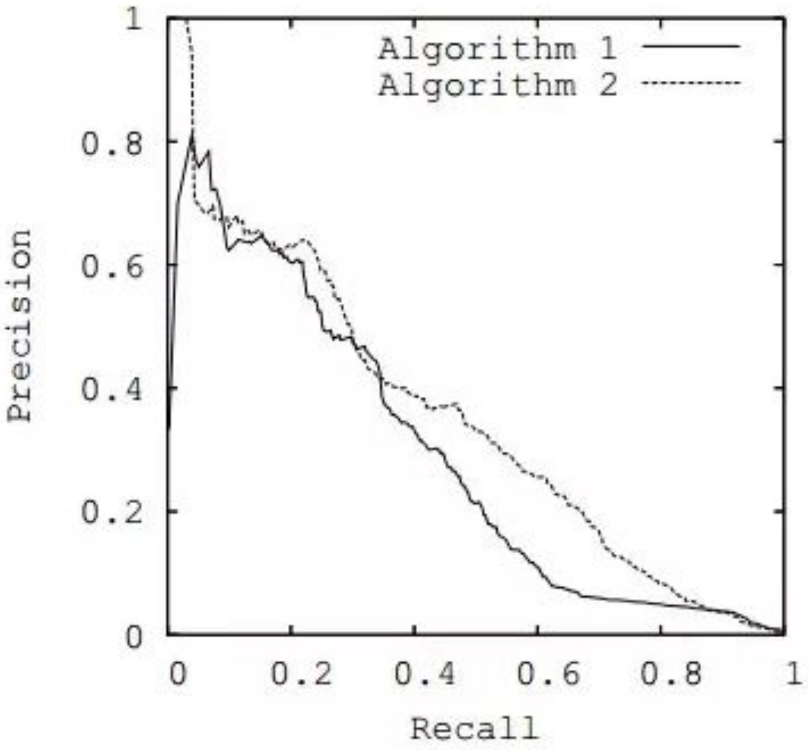

TP
FP
TN
FN
TP
FP
FN
TN


100%
P(TP)
0%
P(FP)
100%

ROC-AUC - area under curve



PR-curve



$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

That’s all. Practice coming next.