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. Extra: once more about regularization.

Linear regression

$$a(x) = \langle w, x \rangle + w_0$$

Linear regression

$$a(x) = \langle w, x \rangle + w_0$$

$$Q = \sum_{i=1}^{N} L\big(y_i, a(x_i)\big)$$

Linear regression

$$a(x) = \langle w, x \rangle + w_0$$

$$Q = \sum_{i=1}^{N} L\big(y_i, a(x_i)\big)$$

$$L\big(y_i, a(x_i)\big) = (y_i - a(x_i))^2$$

Linear regression

$$a(x) = \langle w, x \rangle + w_0$$

$$Q = \sum_{i=1}^{N} L\big(y_i, a(x_i)\big)$$

$$L\big(y_i, a(x_i)\big) = \ (y_i - a(x_i))^2$$

$$L\big(y_i, a(x_i)\big) = |y_i - a(x_i)|$$

Linear classification

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

Linear classification

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

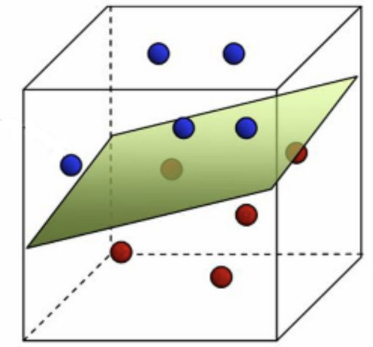
$$f(x) = w_0 + w_1 x_1 + \cdots + w_n x_n$$

Linear classification

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

$$f(x) = w_0 + w_1 x_1 + \cdots + w_n x_n = w_0 + \langle w, x \rangle$$

Geometrical interpretation:
Linearly separable case

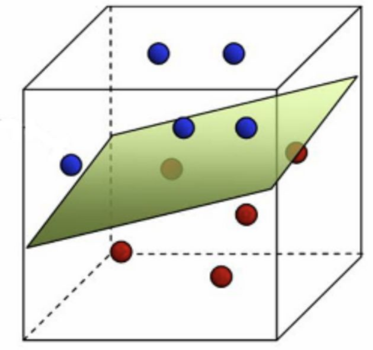


Linear classification

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

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

Geometrical interpretation:
Linearly separable case



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

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

Loss functions

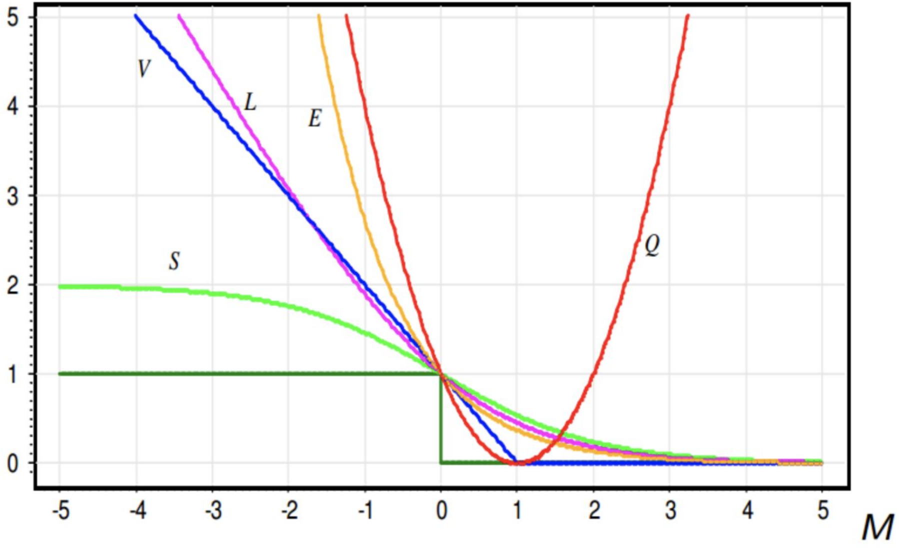
$$Q(w) = \sum_{i=1}^{\ell} [M_i(w) < 0] \leqslant \widetilde{Q}(w) = \sum_{i=1}^{\ell} \mathscr{L}(M_i(w)) \to \min_w;$$

Empirical risk

Loss function

Loss functions

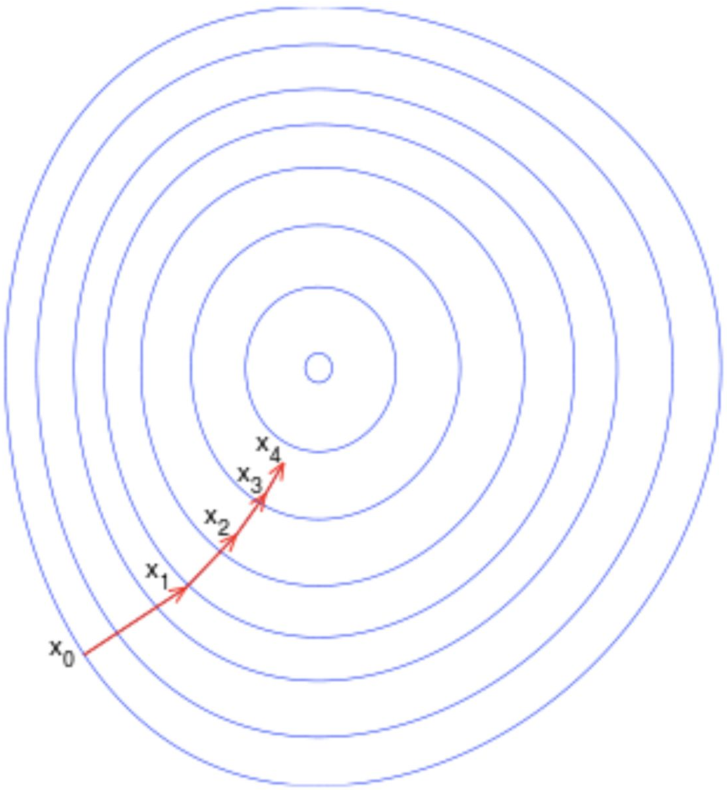
$$Q(w) = \sum_{i=1}^{\ell} \left[M_i(w) < 0\right] \leqslant \widetilde{Q}(w) = \sum_{i=1}^{\ell} \mathscr{L}\left(M_i(w)\right) \to \min_w;$$



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

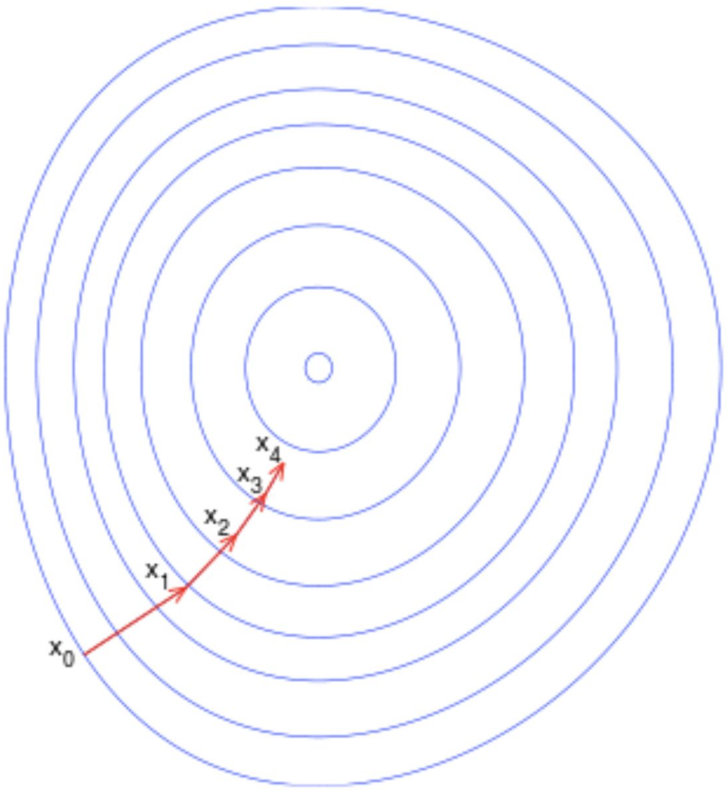
Loss functions

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



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

$$y_i \in \{0, 1\} \qquad Q = -\sum_{i=1}^{\ell} y_i \ln p_i + (1 - y_i) \ln(1 - p_i) \to \min_{w}$$

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

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 regression

$$y_i \in \{0, 1\} \qquad Q = -\sum_{i=1}^{\ell} y_i \ln p_i + (1 - y_i) \ln(1 - p_i) \to \min_w$$

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

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

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

Logistic regression optimization problem

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

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

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

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

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

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

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

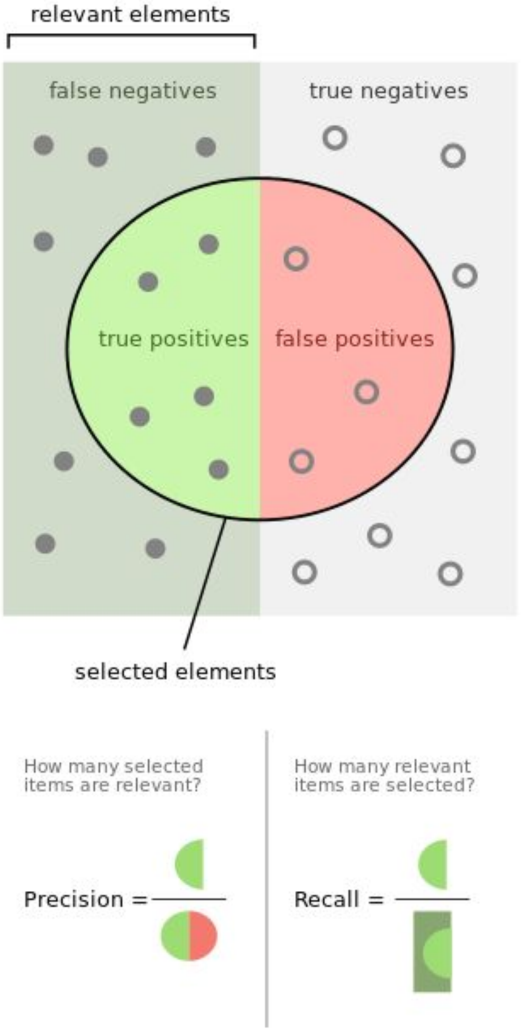
Precision and recall

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

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

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

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.

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

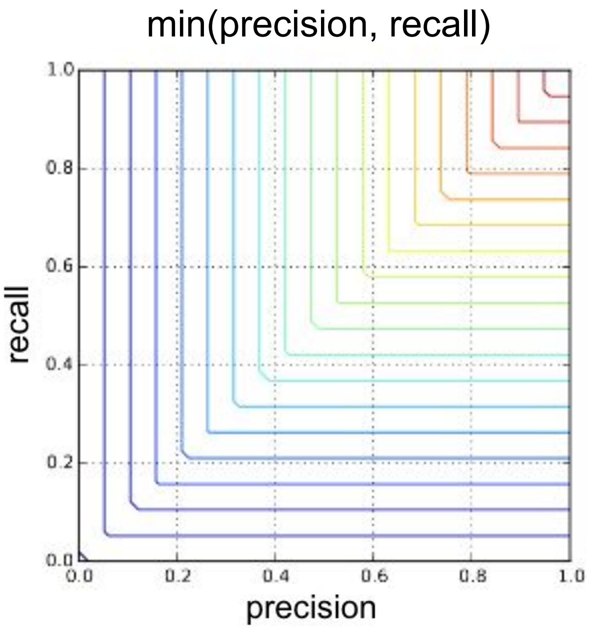
F-score

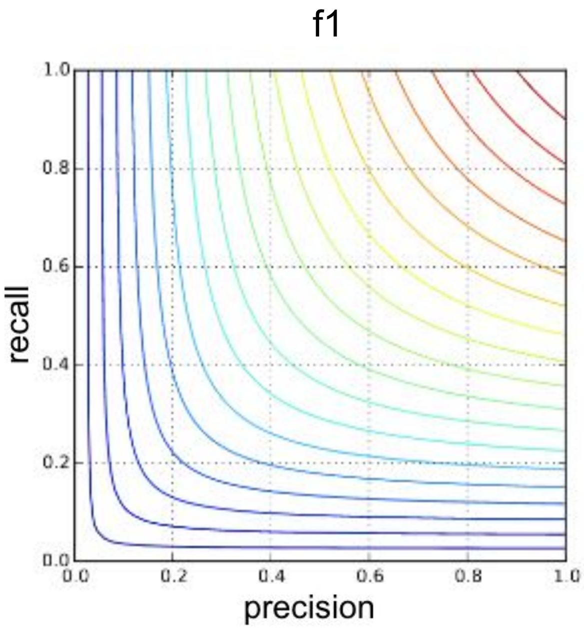
Harmonic mean of precision and recall.
Closer to the smallest one.

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

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

F-score


min(precision, recall)
recall
precision
0.0
0.2
0.4
0.6
0.8
1.0


f1
recall
precision
0.0
0.2
0.4
0.6
0.8
1.0

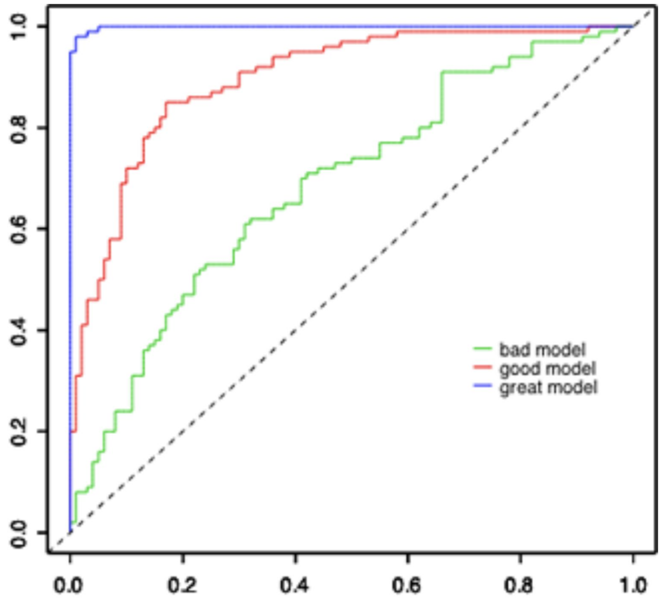
ROC - receiver operating characteristic

		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 - receiver operating characteristic



TPR

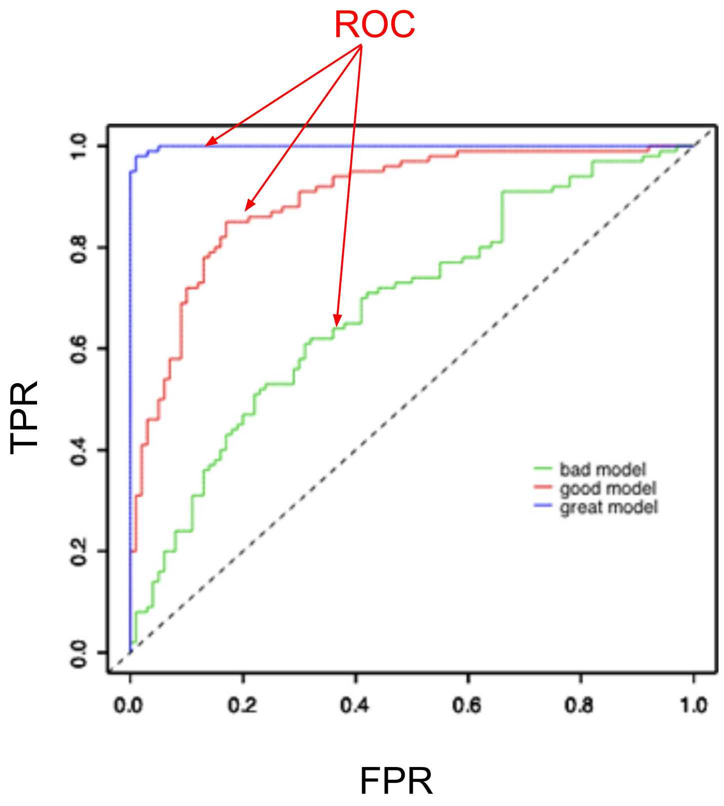
FPR

		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

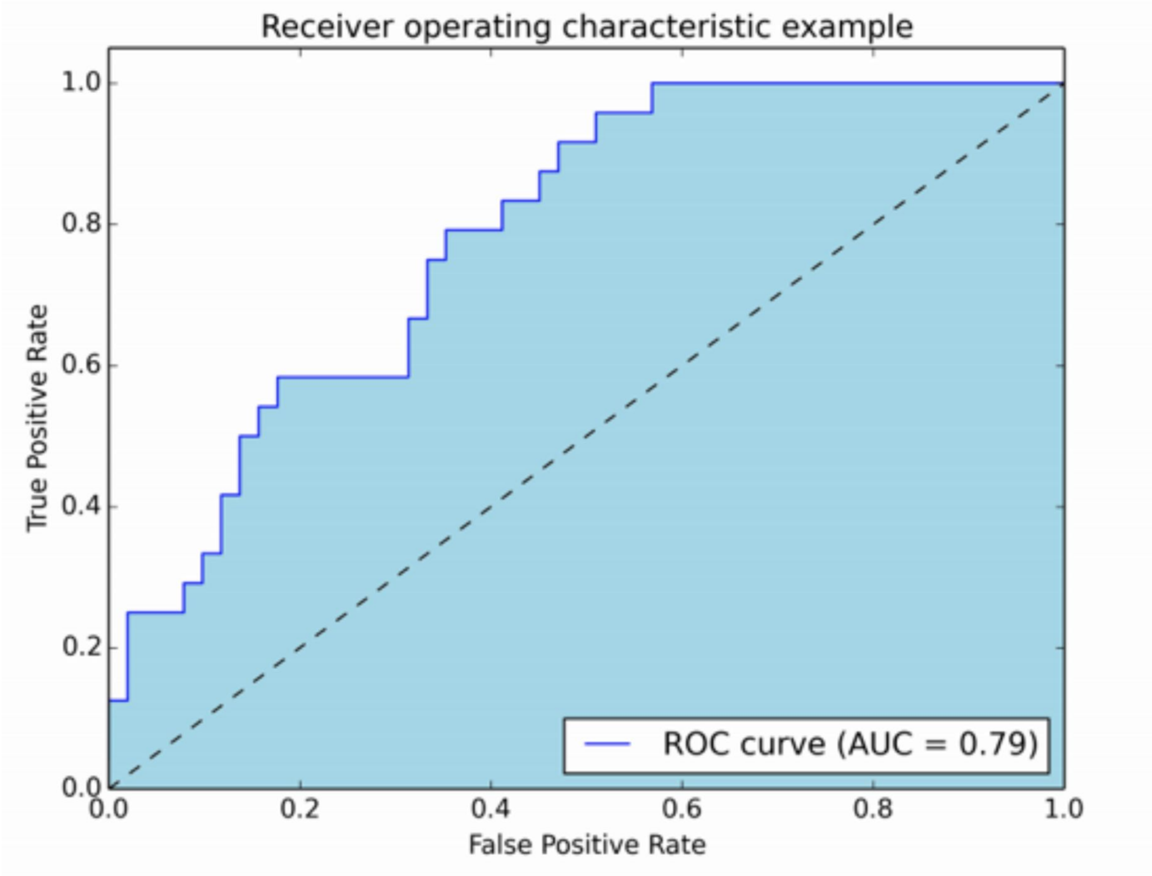


		Actual Class	
		Yes	**No**
Predicted Class	Yes	**T**rue **P**ositive	**F**alse **P**ositive
	No	**F**alse **N**egative	**T**rue **N**egative

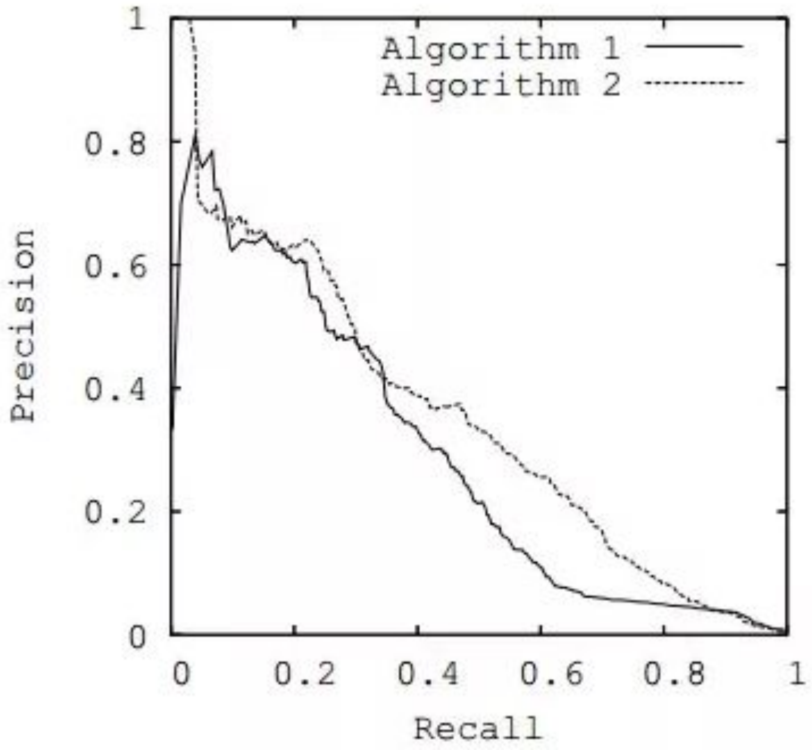
$$TPR = \frac{True\ positives}{True\ positives + False\ negatives}$$

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

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.

Take a look at DMiA group. They provide great courses on Data Mining and Data Analysis.
https://www.facebook.com/groups/data.mining.in.action/
https://vk.com/data_mining_in_action

Extra: once more about regularization

$$Q = \sum_{i=1}^{\ell} L\big(y_i, f(x_i)\big) + \gamma V(w) \rightarrow \min_{w}$$

Extra: once more about regularization

$$Q = \sum_{i=1}^{\ell} L\big(y_i, f(x_i)\big) + \gamma V(w) \to \min_{w}$$

$$\sum_{i=1}^{\ell} -L\big(y_i, f(x_i)\big) - \gamma V(w) \to \max_{w}$$

Extra: once more about regularization

$$Q = \sum_{i=1}^{\ell} L\big(y_i, f(x_i)\big) + \gamma V(w) \to \min_w$$

$$\sum_{i=1}^{\ell} -L\big(y_i, f(x_i)\big) - \gamma V(w) \to \max_w$$

$$\sum_{i=1}^{\ell} \ln e^{-L\left(y_i, f(x_i)\right)} + \ln e^{-\gamma V(w)} \to \max_w$$

Extra: once more about regularization

$$Q = \sum_{i=1}^{\ell} L\big(y_i, f(x_i)\big) + \gamma V(w) \to \min_w$$

$$\sum_{i=1}^{\ell} -L\big(y_i, f(x_i)\big) - \gamma V(w) \to \max_w$$

$$\sum_{i=1}^{\ell} \ln e^{-L\left(y_i, f(x_i)\right)} + \ln e^{-\gamma V(w)} \to \max_w$$

$$e^{-\gamma V(w)} \prod_{i=1}^{\ell} e^{-L\left(y_i, f(x_i)\right)} \to \max_w$$

Extra: once more about regularization

$$Q = \sum_{i=1}^{\ell} L\big(y_i, f(x_i)\big) + \gamma V(w) \to \min_w$$

$$\sum_{i=1}^{\ell} -L\big(y_i, f(x_i)\big) - \gamma V(w) \to \max_w$$

$$\sum_{i=1}^{\ell} \ln e^{-L\left(y_i, f(x_i)\right)} + \ln e^{-\gamma V(w)} \to \max_w$$

$$P(w) \sim \boxed{e^{-\gamma V(w)}} \prod_{i=1}^{\ell} e^{-L\left(y_i, f(x_i)\right)} \to \max_w$$

Extra: once more about regularization

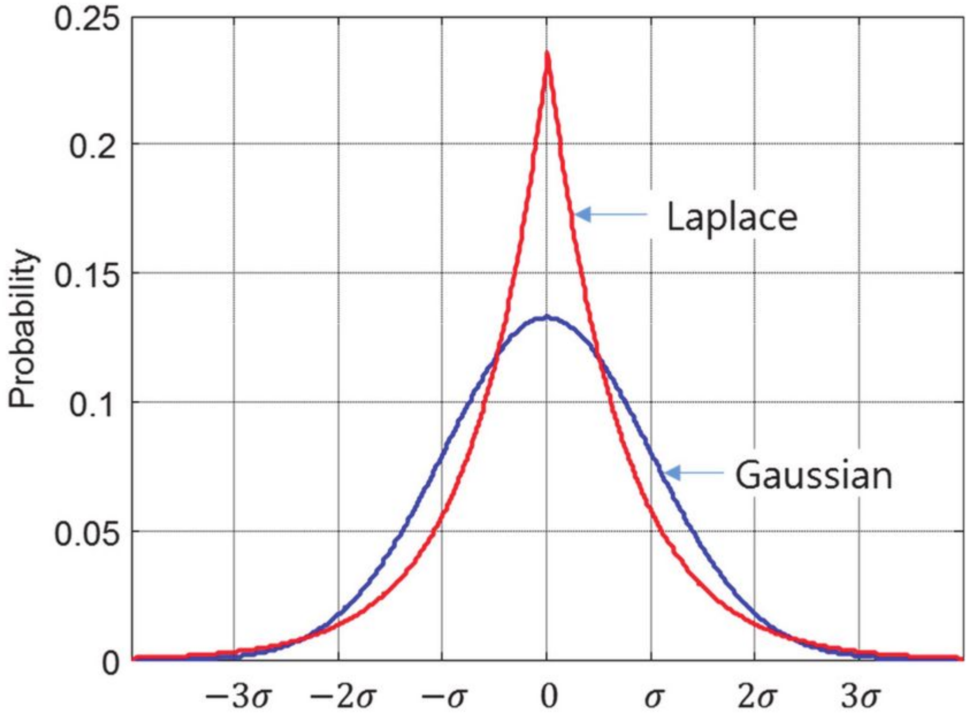
$$Q = \sum_{i=1}^{\ell} L\big(y_i, f(x_i)\big) + \gamma V(w) \to \min_w$$

$$\sum_{i=1}^{\ell} -L\big(y_i, f(x_i)\big) - \gamma V(w) \to \max_w$$

$$\sum_{i=1}^{\ell} \ln e^{-L\left(y_i, f(x_i)\right)} + \ln e^{-\gamma V(w)} \to \max_w$$

$$P(w) \sim \boxed{e^{-\gamma V(w)}} \prod_{i=1}^{\ell} \boxed{e^{-L\left(y_i, f(x_i)\right)}} \sim P(x_i, y_i | w)$$

Extra: once more about regularization



$= L_1$ regularization

$= L_2$ regularization

Extra: hashing trick

L columns

feature
a
b
c
b

hash(a) % L = hash(c) % L = 1	hash(b) % L = 2
1	
	1
1	
	1