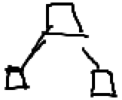
AdaBoost

k_m: 

Dane: (x_i, y_i), $i=1,\dots,N$, $y_i \in \{-1,1\}$

- $C_0(x)=0$
- $C_m(x_i) = C_{m-1}(x_i) + \alpha_m k_m(x_i)$

Jak dobieramy $\alpha_m \in \mathbb{R}$, klasyfikator $k_m : X \to \{-1,1\}$? Tak, żeby zminimalizować funkcję kosztu:

$$E = \sum_{i=1}^{N} e^{-y_i C_m(x_i)} = \sum_{i=1}^{N} \underbrace{e^{-y_i C_{m-1}(x_i)}}_{= w_i^{(m)}} e^{-y_i \alpha_m k_m(x_i)} =$$

$$= \sum_{i: y_i = k_m(x_i)} w_i^{(m)} e^{-\alpha_m} + \sum_{i: y_i \neq k_m(x_i)} w_i^{(m)} e^{\alpha_m} = \sum_{i=1}^{N} w_i^{(m)} e^{-\alpha_m} + \sum_{i: y_i \neq k_m(x_i)} w_i^{(m)} \left(e^{\alpha_m} - e^{-\alpha_m}\right)$$

Dopasowujemy k_m tak, by zminimalizować $\sum_{i: y_i \neq k_m(x_i)} w_i^{(m)}$. Mając ustalone k_m, dobieramy α_m tak, by zminimalizować E.

$$\alpha_m = \frac{1}{2} \ln\left[\left(\sum_{i: y_i = k_m(x_i)} w_i^{(m)}\right) \Big/ \left(\sum_{i: y_i \neq k_m(x_i)} w_i^{(m)}\right)\right].$$

Otrzymujemy C_m, ostateczna klasyfikacja: $\mathrm{sgn}(C_m(x))$

Tabela pomyłek (confusion matrix)

predykowana klasa \ rzeczywista klasa	−	+
−	a_{00}	a_{01}
+	a_{10}	a_{11}

← liczby obserwacji

pred. \ rzeczywiste	−	+
−	TN	FN
+	FP	TP

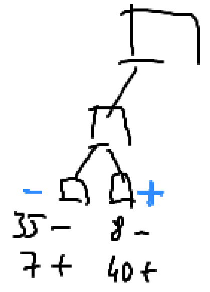
powiedzmy, że tego typu błędu chcemy uniknąć

TN = true negative

FN = false negative

FP = false positive

TP = true -"-



$P(-|x) = \frac{8}{48} = \frac{1}{6}$

Można wprowadzić funkcję kosztu: C

	$C(0,1)$
$C(1,0)$	

np.

	9
1	

$$p^* = \frac{C(1,0)}{C(1,0)+C(0,1)}$$

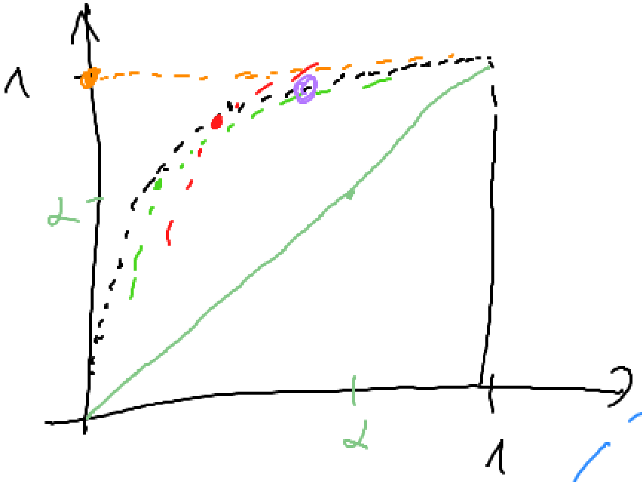
np. $p^* = \frac{1}{1+9} = \frac{1}{10}$

Jeżeli $P(-|x) \geq p^*$, to klasyfikujemy x jako „−"

↑ estymowane prstwo tego, że x jest w klasie −

krzywa ROC (receiver operator characterics)

$$\frac{TP}{TP+FN} = \frac{\alpha\beta N}{\beta N} = \alpha$$



$$\frac{FP}{FP+TN} = \frac{\alpha(1-\beta)N}{(1-\beta)N} = \alpha$$

rzeczywista

przew.	−	+
−	TN	FN
+	FP	TP

suma stała

stała liczba obserwacji

	$(1-\beta)N$ −	$\beta \cdot N$ +
−	$(1-\alpha)(1-\beta)N$	$(1-\alpha)\beta N$
+	$\alpha(1-\beta)N$	$\alpha\beta N$

pole pod krzywą ROC jest pewną miarą jakości klasyfikatora

klasyfikator, który losowo przyporządkowuje do jednej z klas: z p-stwem α : +

$1-\alpha$: −

Klasyfikator bayesowski

przyporządkowujemy obserwację do najbardziej prawdopodobnej klasy, tzn. rozważamy

(x_0, y_0) – obserwacja i jej klasa

$\hat{y}_0$ – przewidywana klasa

$$\hat{y}_0 = \arg\max_k P(Y=k \mid X=x_0)$$

Można pokazać, że błąd $E\left(\sum \frac{1}{N} \mathbb{1}(y_0 \neq \hat{y}_0)\right)$ jest najmniejszy możliwy dla klasyfikatora bayesowskiego

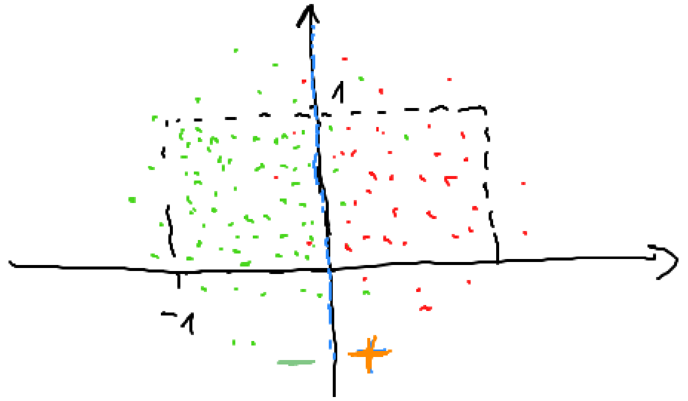
W praktyce nie znamy $P(Y=k \mid X=x_0)$, więc nie możemy użyć tego klasyfikatora.

Np. (x_0, y_0) możemy generować następująco:

$P(y_0=1)=\frac{1}{2}$, $P(y_0=-1)=\frac{1}{2}$ (rzucamy monetą)

Jeśli $y_0=-1$, to $x_0 = u_0 + n$, $u_0 \sim U([-1,0]\times[0,1])$, $n \sim N(0, \sigma^2 I)$, $u_0 \perp n$

$y_0 = 1$, to $x_0 = u_1 + n$, $u_1 \sim U([0,1]\times[0,1])$, $u_1 \perp n$



LDA (Linear Discriminant Analysis) $X \in \mathbb{R}^p$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

$$P(A|B) = \frac{P(A\cap B)}{P(B)}$$

Model

$$P(Y=k\,|\,X=x) = \frac{\pi_k \cdot f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)},$$

gdzie $\pi_k = P(Y=k)$ — łatwe do estymacji

$f_k(x) =$ „$P(X=x\,|\,Y=k)$" funkcja gęstości

$$P(X\in A\,|\,Y=k) = \int_A f_k(x)\,dx$$

Żeby poradzić sobie z f_k, zakładamy, że f_k ma pewną konkretną postać

$p=1$

$$f_k(x) = \frac{1}{\sqrt{2\pi}\,\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right),$$

μ_k, σ_k^2 — być może różne dla różnych klas k

LDA: $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_K^2 = \sigma^2$ (QDA: jak LDA, ale bez tego założenia)

Wtedy

$$P(Y=k\mid X=x) = \frac{\pi_k \cdot \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K}\pi_l \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

Przyporządkujemy x klasę k, dla której powyższe wyrażenie jest największe

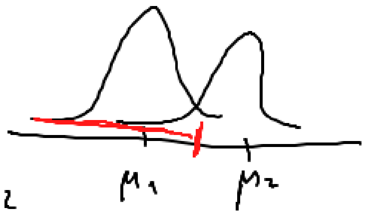
Mianowniki są takie same dla każdego $k=1,\dots,K$, więc wystarczy szukać największego licznika, lub jego logarytmu:

$$\log\pi_k + \underbrace{\log\frac{1}{\sqrt{2\pi}\,\sigma}}_{\text{stałe}} - \frac{1}{2\sigma^2}(x-\mu_k)^2 = \log\pi_k + \log\frac{1}{\sqrt{2\pi}\,\sigma} - \frac{1}{2\sigma^2}x^2 + \frac{1}{\sigma^2}x\mu_k - \frac{1}{2\sigma^2}\mu_k^2$$

Gdy $K=2$: $$\log\pi_1 + \frac{1}{\sigma^2}x\mu_1 - \frac{1}{2\sigma^2}\mu_1^2 \overset{?}{>} \log\pi_2 + \frac{1}{\sigma^2}x\mu_2 - \frac{1}{2\sigma^2}\mu_2^2 \quad \Big|\cdot\sigma^2$$

$\pi_1=\pi_2$: $$(\mu_1-\mu_2)x > \frac{1}{2}(-\mu_2^2+\mu_1^2) = \frac{1}{2}(\mu_1-\mu_2)(\mu_1+\mu_2)$$

np.



$\mu_1<\mu_2$

jeśli $x < \frac{1}{2}(\mu_1+\mu_2)$ to x jest klasyfikowany do klasy 1