$$\mathbb{P}\left(\{y^{(i)}\} \mid \{\vec{x}^{(i)}\}, \vec{\beta}\right) = \prod_{i=1}^{N} h_{\vec{\beta}}(\vec{x}^{(i)})^{y^{(i)}} \left(1 - h_{\vec{\beta}}(\vec{x}^{(i)})\right)^{1-y^{(i)}}$$

log-likelihood:

$$\mathcal{L} = \sum_{i=1}^{N} y^{(i)} \log h_{\vec{\beta}}(\vec{x}^{(i)}) + (1-y^{(i)}) \log\left(1 - h_{\vec{\beta}}(\vec{x}^{(i)})\right)$$

$\Rightarrow \nabla_{\vec{\beta}} \mathcal{L}$ to be maximized.

Our model.

$$\begin{cases} P(y=1) = h_{\vec{\beta}}(\vec{x}^{(i)}) = \dfrac{1}{1+e^{-\vec{x}^{(i)T}\cdot\vec{\beta}}} = h(\vec{x}^{(i)T}\cdot\vec{\beta}) \\[2ex] \mathbb{P}(y=0) = 1 - h_{\vec{\beta}}(\vec{x}^{(i)}) = \dfrac{e^{-\vec{x}^{(i)T}\cdot\vec{\beta}}}{1+e^{-\vec{x}^{(i)T}\cdot\vec{\beta}}} = e^{-\vec{x}^{(i)T}\cdot\vec{\beta}}\, h(\vec{x}^{(i)T}\cdot\vec{\beta}) \end{cases}$$

Note that: $\log\left(1 - h_{\vec{\beta}}(\vec{x}^{(i)})\right) = -\vec{x}^{(i)T}\cdot\vec{\beta} + \log\left(h(\vec{x}^{(i)T}\cdot\vec{\beta})\right)$

$$\mathcal{L} = \sum_{i=1}^{N} \underbrace{\log h_{\vec{\beta}}(\vec{x}^{(i)})} - (1-y^{(i)})(\vec{x}^{(i)T}\cdot\vec{\beta})$$

$$\nabla_{\vec{\beta}}\left[\log h_{\vec{\beta}}(\vec{x}^{(i)})\right] = \frac{e^{-\vec{x}^{(i)T}\cdot\vec{\beta}}}{1+e^{-\vec{x}^{(i)T}\cdot\vec{\beta}}}\, \vec{x}^{(i)}$$

$$= \left(1 - h_{\vec{\beta}}(\vec{x}^{(i)})\right) \vec{x}^{(i)}_{(p+1)\times 1}$$

$$\nabla_{\vec{\beta}} \mathcal{L} = \sum_{i=1}^{N} \left[1 - h_{\vec{\beta}}(\vec{x}^{(i)})\right] \vec{x}^{(i)} - (1 - y^{(i)}) \vec{x}^{(i)}$$

$$= \sum_{i=1}^{N} \left[y^{(i)} - h_{\vec{\beta}}(\vec{x}^{(i)}) \right] \vec{x}^{(i)}$$

Maximum:

$$\nabla_{\vec{\beta}} \mathcal{L} \Big|_{\vec{\beta}^*} = 0 \qquad \boxed{\sum_{i=1}^{N} \left[y^{(i)} - h(\vec{x}^{(i)T} \cdot \vec{\beta}^*_{log}) \right] \vec{x}^{(i)} = 0_v}$$

Equivalent of normal equations in LS.

(p+1) equations

$\mathcal{L}$ is concave $\Rightarrow$ ~~a~~ global maximum is $\vec{\beta}^*_{log}$

which can be found by standard optimisation.

$$P(\hat{y}=1) = h(\vec{x}^{in T} \cdot \vec{\beta}^*_{log}) \in [0, 1]$$

$$\boxed{\text{classifier:} \quad P(\hat{y}=1 \mid \vec{x}^{in}) > \tau \quad \vec{x}^{in} \in \{1\}}$$

τ is the threshold for the classifier: usually $\tau = \frac{1}{2}$.

[But it can be tuned a posteriori]

Summary of our solution for logistic regression:

$$X_{N \times (p+1)} \qquad X\vec{\beta}^* = \vec{\hat{y}}_{N \times 1}$$

logistic regression

Normal equations

$$\boxed{X^T\left[\vec{y} - \vec{h}(X\vec{\beta}^*_{log})\right] = 0}$$

$$\vec{h}_{N \times 1}(X\vec{\beta}^*) \Rightarrow h_i = \left(\frac{1}{1+e^{-\vec{x}^{(i)T}\vec{\beta}^*_{log}}}\right)$$

log

compare

Least squares

Remember that for LS:

Normal equations for LS

$$\boxed{X^T\left[\vec{y} - X\vec{\beta}^*_{LS}\right] = 0}$$

Quality of the classifier.

confusion matrix

or

contingency table

Two-classes

For a $\{0,1\}$ classifier

		True: 1	True: 0
Predicted	1	True positives (TP)	False positives (FP)
	0	False negatives (FN)	True negatives (TN)

Type I error

Type II errors

We want most cases on the diagonal

True positive rate $= \dfrac{TP}{TP+FN} = TPR$
$\parallel$
Sensitivity or recall

True negative rate : $\dfrac{TN}{TN+FP} = TNR$
$\parallel$
Specificity

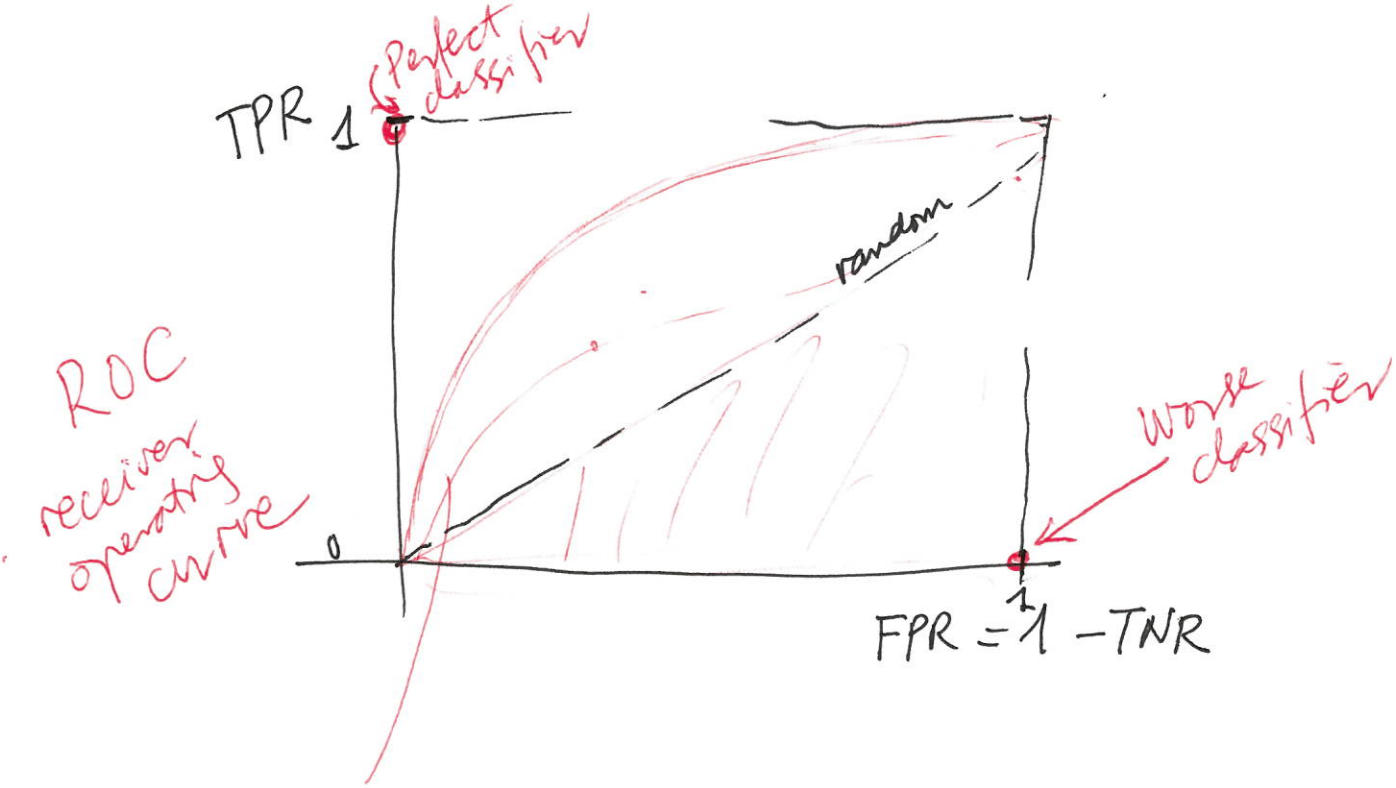
Accuracy : $\dfrac{TP+TN}{N_{validation}}$.

Precision $= \dfrac{TP}{TP+FP}$

Recall $= \dfrac{TP}{TP+FN}$

Precision $= \dfrac{TP}{TP+FP}$

$$\boxed{F = \frac{2\ \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}}$$



AUC : Area Under the Curve is an overall measure of quality

$\boxed{AUC > \frac{1}{2}}$ Better than random over a range of parameters