Classification Assessment

$(\vec{x}_i, y_i)_{i=1}^{n}$

$\vec{x} \in \mathbb{R}^d$

$y_i \in \{1, 2, \ldots, k\}$ # of classes

Confusion Matrix (k×k matrix)

True $k=3$

Predictions

	c_1	c_2	c_3	
c_1	n_{11}	n_{12}	n_{13}	m_1
c_2	n_{21}	n_{22}	n_{23}	m_2
c_3	n_{31}	n_{32}	n_{33}	m_3
	n_1	n_2	n_3	

$M(\vec{z}) = \hat{y}$

model, prediction

$n_{ij} = \#\{\hat{y} = c_i,\ \text{but}\ y = c_j\}$

diagonal entries ⇒ accuracy

off-diagonal ⇒ errors

n_i = # of test cases belonging to class i

m_i = # of cases the classifier predicts to be in c_i

Precision/accuracy for c_i

$$p_i = \frac{n_{ii}}{m_i}$$

# of correct predictions for c_i

# of predictions for c_i

recall/coverage for c_i

$$r_i = \frac{n_{ii}}{n_i}$$

# of correct predictions

of "true" instance of c_i

Precision-recall tradeoff

F-score for c_i

$$f_i = \frac{2}{\frac{1}{p_i} + \frac{1}{r_i}} = \frac{2 \cdot p_i \cdot r_i}{p_i + r_i}$$

Overall k-Classes:

$$\text{avg F-score} = \frac{1}{k} \sum_{i=1}^{k} f_i$$

K=2

		True P_1	N_2	
Prediction	P_1	TP n_{11}	FP n_{12}	m_1
	N_2	FN n_{21}	TN n_{22}	m_2
		n_1	n_2	

TPR : True Positive Rate (Sensitivity)

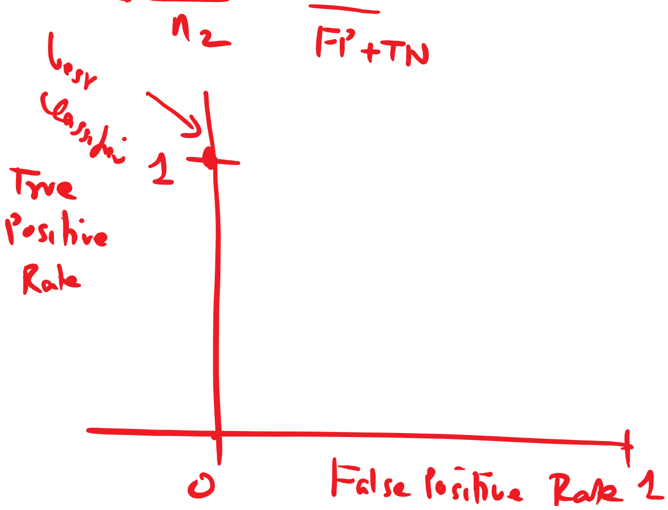
$$\frac{TP}{n_1} = \frac{TP}{TP + FN}$$

recall for P

FPR : False Positive Rate (1 − Specificity)

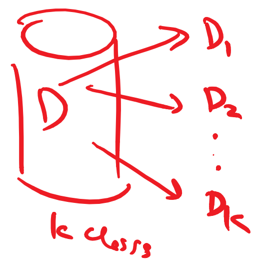
$$= \frac{FP}{n_2} = \frac{FP}{FP + TN}$$

TNR



SVM for k-classes

One-versus-rest approach

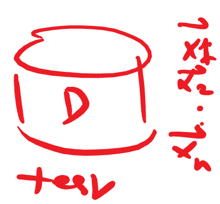


$$\underbrace{\frac{D_i}{P} \quad \text{vs.} \quad \frac{\bigcup_{j \neq i} D_j}{N}}$$

k-SVM

k-hyperplanes

(FPR, TPR)



0 False Positive Rate 1

test $\vec{x}_n$

n test case

ROC Curve → Receiver Operating Curve

Naive Bayes

$NB(\vec{x}_i) \rightarrow P(C_i = Pos \text{ or } +1)$

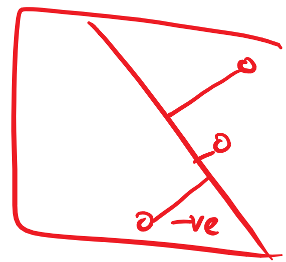
Pos

$X_1 \rightarrow P(X_1) \rightarrow S_1$

$X_2 \rightarrow P(X_2) \rightarrow S_2$

$\vdots$

$X_n \rightarrow P(X_n) \rightarrow S_3$



Scores
→ Probabilities (NB)
→ distance (SVM)
signed

① Sort the score from the best to the worst

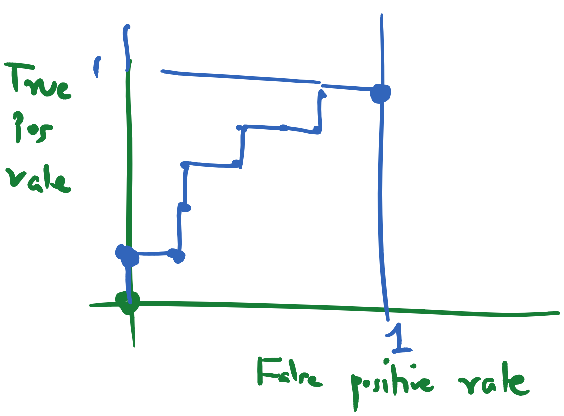
true P N

S_1 S_2 S_1 S_5 S_7 S_8 S_3 S_4 S_6 S_9 → low

Predict P P N

② Choose <u>all</u> possible thresholds

Initially: No Positive, all points are negative



$TP = 0$

$FP = 0$

(FPR, TPR)

$= (0, 0)$

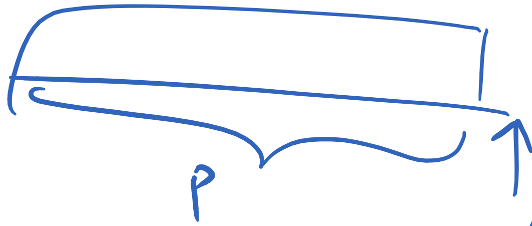
$(1/n_1, 0)$

$(1/n_1, 1/n_2)$

	n_1	n_2
P	1_{TP}	1_{FP}
N		

$(1/n_1, 1/n_2)$

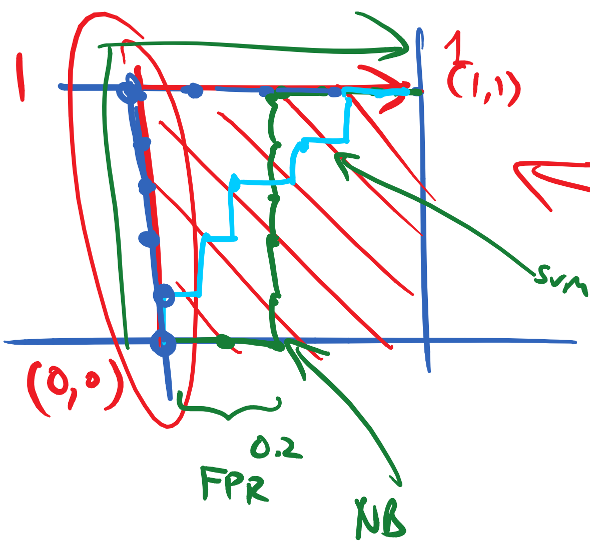
finally



no negative

$(FPR, TPR) = (1,1)$

n_1	n_2
0	0
n_1	n_2

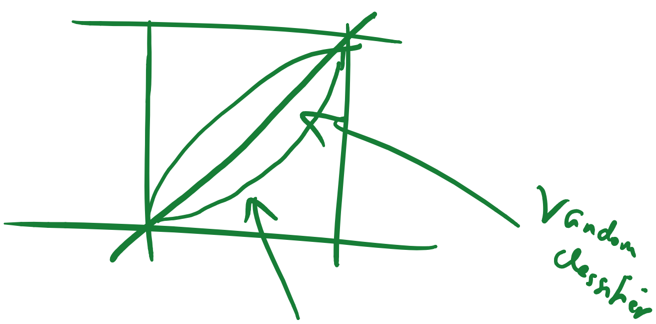


best/optimal/ideal classifier ROC curve

Area Under the curve

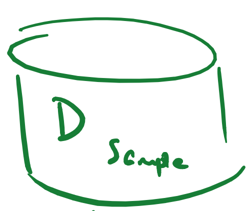
AUC = 1

higher AUC better classifier



← AUC: 0.5

SVM:



D_{train} 80% $\qquad$ D_{test} 20%

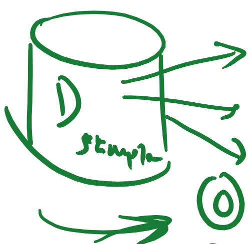
SVM

Use D_{test} on SVM $\longrightarrow$ 1 number (global accuracy)

① K-fold cross validation

5-fold

r Iterations



D_1, ⋮, D_K } folds

⓪ random shuffle

① split into k equal parts

	test	train
$a_1 \leftarrow$	D_1	$D_2 \dots D_5 \rightarrow$
$a_2 \leftarrow$	D_2	$D_1, D_3 \dots D_5$
	⋮	
$a_5 \leftarrow$	D_5	$D_1 \dots D_4$

K-different performance numbers

0.8, 0.85, 0.9, 0.75, 0.8

a_1 a_2 a_3 a_4 a_5

over all $r \cdot K$ value

$$\hat{\mu} = \frac{1}{K} \sum_{i=1}^{K} a_i$$

estimated mean accuracy

$$\hat{\sigma}^2 = \frac{1}{K} \sum_{i=1}^{K} (a_i - \hat{\mu})^2$$

Variance

② LOOCV

Leave one out cross-validation

$|D_{sample}| = n$ n-fold

test	rest of data train
$\vec{x}_1$	
x_2	
x_3	
⋮	
$\vec{x}_n$	

Confidence interval for the true mean accuracy?

$$\underbrace{\theta_1 \quad \theta_2 \ldots \quad \theta_k}_{k \text{ values}}$$

θ : accuracy

$$\boxed{\hat{\mu} = \frac{1}{k}\sum_{i=1}^{k}\theta_i}$$

estimated mean score/accuracy

$$\boxed{\hat{\sigma}^2 = \frac{1}{k}\sum_{i=1}^{k}(\theta_i - \hat{\mu})^2}$$

$\mu \leftarrow$ true mean!

$$Z_k = \frac{\hat{\mu} - \mu}{\sqrt{Var(\hat{\mu})}} = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{k}}$$

Z-score variable

how much does the mean vary?

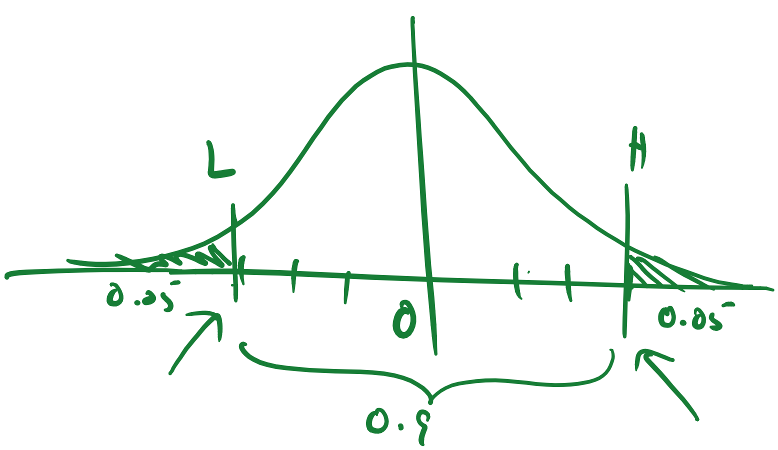
$$\boxed{Var(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{k}}}$$

Standard error of the mean.

Standard Normal distribution

Confidence bound on true mean!

Confidence value = 0.9



90% of the time $L \leq Z_k \leq H$

90% confidence interval

$$L \le \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{k}} \le H$$

$$\frac{\hat{\sigma}}{\sqrt{k}} L \le \hat{\mu} - \mu \le H \cdot \hat{\sigma}/\sqrt{k}$$

$$\left(\underbrace{\underline{\underline{\hat{\mu} - \frac{\hat{\sigma}}{\sqrt{k}} L}} \ge \mu \ge \underline{\underline{\hat{\mu} - H \cdot \hat{\sigma}/\sqrt{k}}}} \right)$$

we can now claim that "true" mean lies in this interval with 90% Confidence

---

In practice k is small

$$Z_k = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{k}}$$

follows a t-distribution with $k-1$ degrees of freedom

Instead of normal use t-distribution

↓

effect of $\underline{L}$ & $\underline{H}$

t-distribution converge to Normal as $k \to \infty$ (30)

$k=5$

Comparing multiple classifiers

Comparing multiple classifiers

SVM vs. NN vs NB . .

$\hat{\mu}$ accuracy 0.95   0.97 $\hat{\mu}$ accuracy   0.98 $\hat{u}$ accuracy

$\hat{\sigma}$ 0.1   0.1   0.1

Is the difference "statistically significant"

Paired t-test

k-fold CV. M_A vs M_B

on the same folds
$$\begin{cases} \theta_1^A \ \theta_2^A \ \ldots \ \theta_k^A \\ \theta_1^B \ \theta_2^B \ \ldots \ \theta_k^B \end{cases}$$

$\delta_i : \theta_i^A - \theta_i^B \leftarrow$ diff in performance of A & B

$\delta_1 \quad \delta_2 \ \ldots \ \delta_k \rightarrow \hat{\mu}_\delta, \ \hat{\sigma}_\delta$

NULL hypothesis

$\mu_\delta = 0$

diff

Alternative

$\mu_\delta \neq 0$

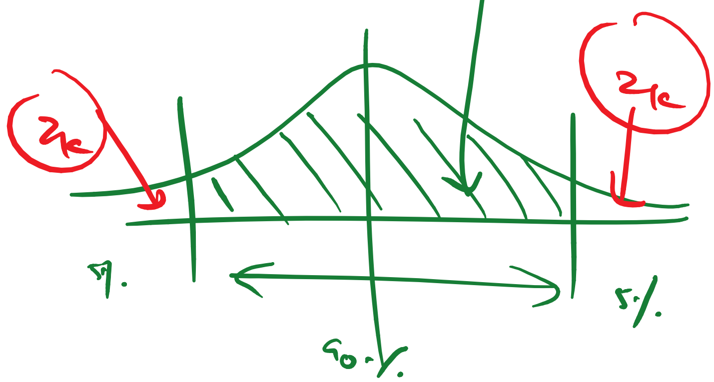
$$Z_k = \frac{\hat{\mu}_\delta - \mu_\delta}{\hat{\sigma}_\delta / \sqrt{k}}$$

$$Z_k = \frac{\hat{\mu}_\delta - \mu_\delta}{\hat{\sigma}_\delta / \sqrt{k}}$$

$$Z_k = \frac{\hat{\mu}_\delta}{\hat{\sigma}_\delta / \sqrt{k}}$$

t - distribution

90% confidence



If Z_k lies in 90% interval

we cannot reject H_0

$\Rightarrow$ A & B have

statistically same

performance

else

Z_k lie outside

$\Rightarrow$ A & B <u>are</u>

statistically significantly

different