Lecture 17

Sunday, October 29, 2023 9:38 PM

Exam II (Thursday, Nov 2)

1) Logistic regression (Chap 24) — log odd ratio
2) Neural Networks (Chap 25) — net gradients, ∇_w, ∇_b, output activations
3) Bayes classifier (Chap 18) — naive, full (no categorical data)
   ↳ KNN classifier
4) Decision Trees (Chap 19) — split eval, metrics
5) Support Vector Machines (Chap 21) — a) margin vs. slacks
   b) given $\vec{\alpha}$, compute $\vec{\omega}$ (linear)
   ↳ polynomial (quadratic)

$$\vec{x} \rightarrow \underbrace{\phi(\vec{x})}_{\text{feature space}}$$

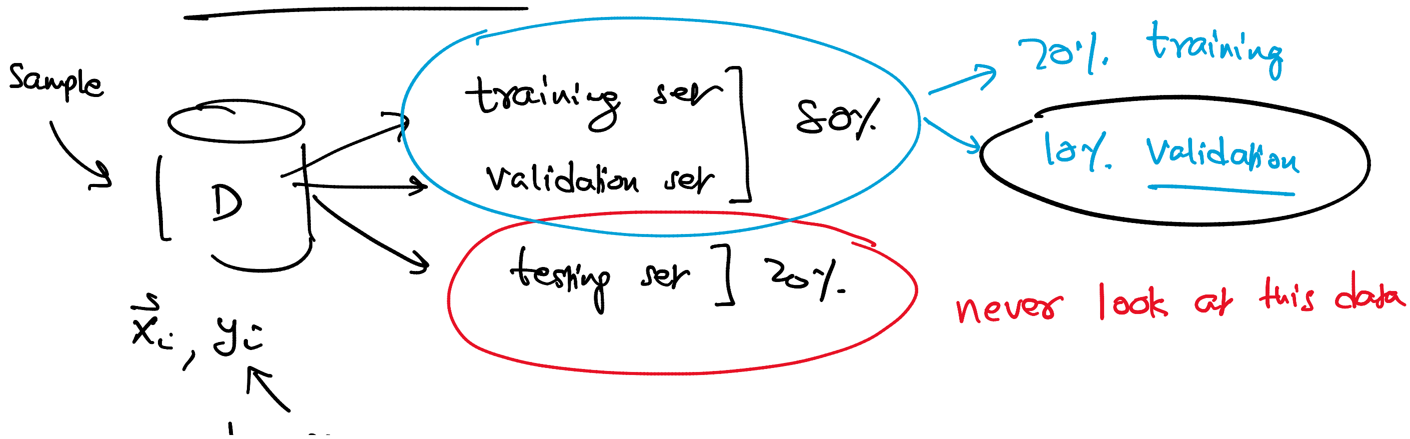
3

$$\vec{w} = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i)$$

   c) Some iterations of SGA (find $\vec{\alpha}$)

6) Classification assessment (see 22.1 only)
   ↳ Prec, recall, f1, Roc (TPR vs FPR)

Classification Assessment



x, y

true response

User-specified constraints

↳ hyperparameters

↓

try diff C values (grid search) ← C value in SVMs

learning rate

$[10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4} \ldots]$

build model on training

eval on validation

Once we have a trained & validated model

↳ then apply on test data

↳ compute f1-score

---

k-fold cross validation (CV)

↳ repeat k time : build k different models

↳ compute $\underline{\theta_i}$ $i = 1 \ldots k$ (on the test data)

some measure, e.g. accuracy, f1,

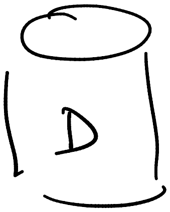
↳ $\hat{\mu} = \text{mean}\{\theta_1, \theta_2, \ldots, \theta_k\}$

↳ $\hat{\sigma}^2 = \text{Var}\{\theta_1, \theta_2, \ldots, \theta_k\}$

0.85

0.15

$0.85 \pm \sqrt{0.15}$

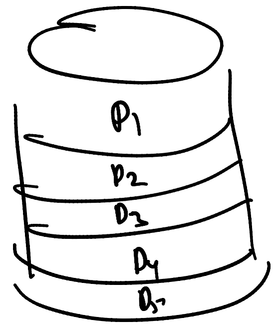


Sample

$k=5$

5-fold CV

1) shuffle D
2) create 5 equal partitions



5 – folds

3) keep aside D_i for testing

Use remaining folds for training & validation

$D \setminus D_i = D - D_i$

$T_i = D \setminus D_i \leftarrow$ training

$D_i \leftarrow$ testing

$\theta_i \leftarrow$ Model (D_i)

test	train
D_1	$D_2\ D_3\ D_4\ D_5$
D_2	$D_1\ D_3\ D_4\ D_5$
$\vdots$	
D_5	$D_1\ D_2\ D_3\ D_4$

LOOCV : leave one out CV

↳ usually used for data which is expensive to collect

n-fold CV

1 point for testing

n-1 points for training

1 point for testing $\quad$ $n-1$ points for training

Bootstrap sampling

sample random dataset of size n $\quad$ D_i $\quad$ $\forall\, i = 1..k$

→ Sample with replacement

each point can be selected multiple times

→ training set : D_i (further split into validation)

→ test on D

$D_i \subseteq D$

θ_i ← over-estimate (optimistic estimate)

since we have seen many of the test points in D_i

Q: sample with replacement, n trials

what is the probability that a point will not be sampled?

a) what is the prob that a point is sampled

$\frac{1}{n}$

b) prob that it is not sampled

$\left(1 - \frac{1}{n}\right)$ for one trial

c) not sampled even after n trials

$$P(x_j \notin D_i) = \left(1 - \frac{1}{n}\right)^n \simeq e^{-1} = 0.368$$

$\underbrace{P(x_j \notin D_i)}$ a point is not sampled for D_i

D_i contains only about 63.2% of the points!

From $\underbrace{\hat{\mu} \;\&\; \hat{\sigma}^2}_{\text{estimates}}$ (mean & variance) for a classifier

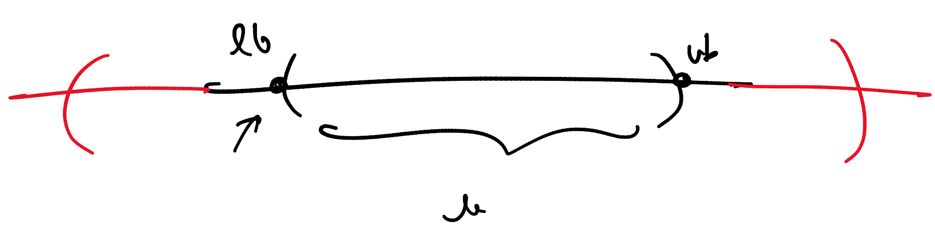
we get a confidence bound on the true mean

expected performance (accuracy, f1)

μ ← true mean.

Compute the confidence interval for the true mean

$$\left(\underbrace{\hat{\mu} - t_k^{\alpha}(\hat{\sigma})}_{lb} \leq \mu \leq \underbrace{\hat{\mu} + t_k^{\alpha}(\hat{\sigma})}_{ub} \right)$$



α : confidence level 95%

t_k : student's t-distribution

α : confidence level

90%

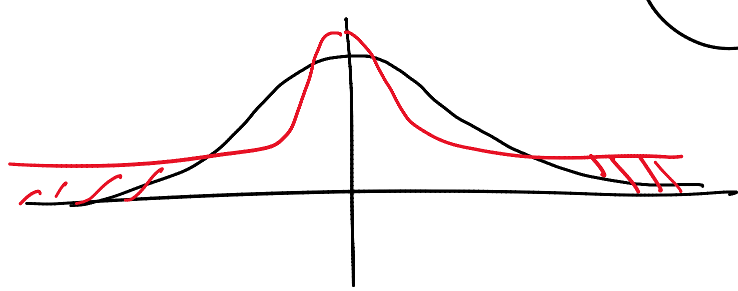
$\alpha \Leftarrow 99\%$

t_k : student's t-distribution with k-1 degrees of freedom

t_k is like the normal as $k \longrightarrow \infty$

t_k is small sample version of normal

k is small e.g. $k=5$, $k=10$

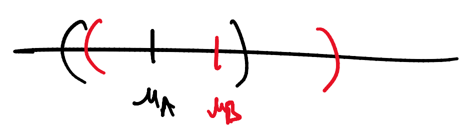


Q How to compare two models?

M_A & M_B

(SVM) (logistic regression)

$\hat{\mu}_A \pm \hat{\sigma}_A$ vs $\hat{\mu}_B \pm \hat{\sigma}_B$



Paired t-test

CV k-folds

for i = 1 . . k

$T_i = D \setminus D_i \longleftarrow$ training

$T_i = D \setminus D_i$ ← training

$D_i =$ i-th fold ← testing

train M_A & M_B on T_i

accuracy A, B

$\theta_i^A = M_A(\underline{D_i})$

$\theta_i^B = M_B(\underline{D_i})$

$\delta_i = \theta_i^A - \theta_i^B$: difference in performance

$\hat{\mu}_\delta \equiv$ mean difference $=$ mean$\{\delta_1, \delta_2, \dots, \delta_k\}$

$\hat{\sigma}_\delta^2 \equiv$ Variance of difference

Hypothesis testing

Q. Are the two models different in terms of θ (accuracy)

H_0: null hypothesis

a) there is no difference between the two models

b) look for evidence to support or reject H_0

H_a: alternative hypothesis

→ there is a difference

H_0: there is no difference

$\mu_\delta = 0$ true expected difference

$\mu_\delta = 0$ true expected difference

t-test:

predicted mean

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

true mean

std of the mean

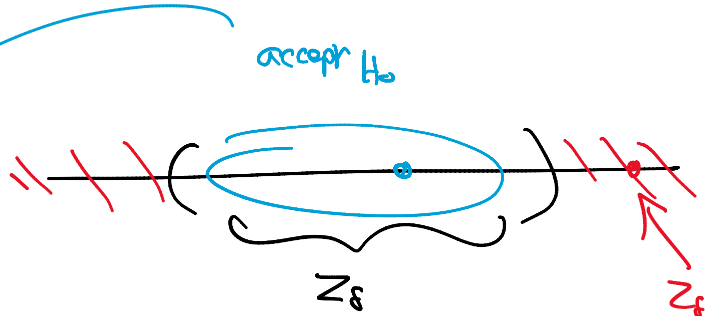
standardized diff score (how many deviations away from true mean)

this has a t-distribution with $k-1$ degrees of freedom

$$\underbrace{\left(-t^k_{\alpha/2} \le Z_\delta \le +t^k_{\alpha/2}\right)}_{\text{what is expected under the null hypothesis}}$$

α : 90%

α : 98%



If Z_δ lies outside the confidence interval then reject H_0

$\Rightarrow$ there is a statistically significant difference

If Z_δ lies in the confidence interval

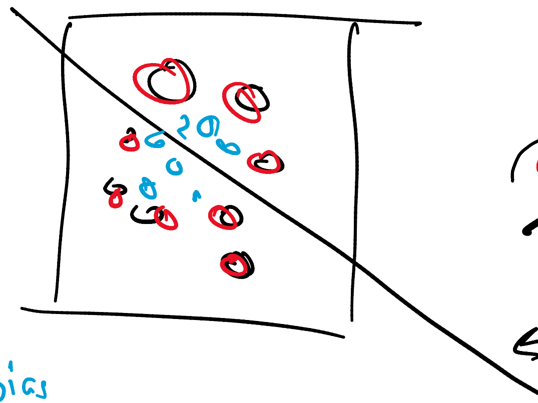
$\Rightarrow$ H_0 is probably the right conclusion

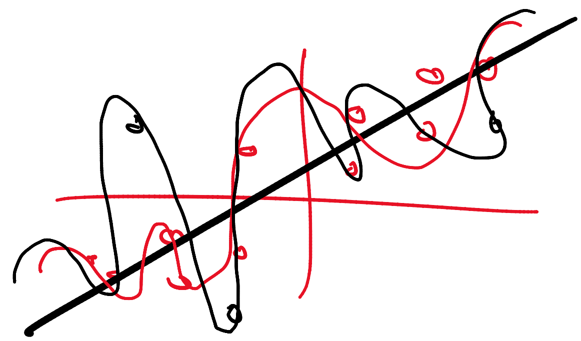
⇒ H_0 is probably the right conclusion

⇒ no difference!

Bias-Variance of the classifier

Bias ← Inherent class of functions a model can approximate.





⟷ model = linear

bias			variance
linear models	→	**high bias**	→ low variance
NN (Complex functions)	→	**low bias**	→ **high variance**

error = bias + variance

bias-variance tradeoff