Umetna inteligenca

Osnovni principi strojnega učenja

1. učenje kot modeliranje,
2. princip najkrajšega opisa
3. inkrementalno učenje,
4. princip večkratne razlage,
5. ocenjevanje verjetnosti

Učenje kot modeliranje

- Učenje je opisovanje oz. modeliranje podatkov
- Učni in izvajalni algoritem
- Učni podatki = opisi problemov in njihovih rešitev
- Novi podatki = opisi novih, nerešenih problemov
- Predznanje = prostor možnih modelov + kriterij optimalnosti + začetna hipoteza + množica hevristik + …

Strojno učenje = optimizacija:

- **Dano:** prostor možnih rešitev + kriterijska funkcija
- **Poišči:** rešitev, ki optimizira kriterijsko funkcijo

Princip najkrajšega opisa (MDL)

It is vain to do with more what can be done with fewer.

William of Ockham

- hipoteza, ki čim bolj ustreza vhodnim podatkom in predznanju
- Princip *Ockhamove britve* (Occam's Razor Principle):

Najpreprostejša razlaga je najbolj zanesljiva (verjetna).

Kriteriji kvalitete hipoteze:

- maksimizirati napovedno točnost hipoteze,
- minimizirati povprečno ceno napak,
- minimizirati velikost hipoteze,
- maksimizirati prileganje hipoteze vhodnim podatkom,
- maksimizirati razumljivost hipoteze,
- minimizirati časovno zahtevnost napovedovanja,
- minimizirati število parametrov, potrebnih za napovedovanje,
- minimizirati ceno pridobivanja vrednosti parametrov,
- **maksimizirati verjetnost hipoteze**

Princip najkrajšega opisa (MDL)

$\mathcal{H}$ - množica možnih hipotez

$H \in \mathcal{H}$ - hipoteza,

B - predznanje

E - vhodni podatki

Optimalna hipoteza:

$$H_{opt} = arg \max_{H \in \mathcal{H}} P(H|E, B)$$

Princip najkrajšega opisa (MDL)

$P(H|B)$ - (apriorna) verjetnost hipoteze

Apriorna količina informacije hipoteze H:

$$I(H|B) = -\log_2 P(H|B) \quad [bit]$$

Aposteriorna količina informacije hipoteze H:

$$I(H|E,B) = -\log_2 P(H|E,B) \quad [bit]$$

Optimalna hipoteza:

$$H_{opt} = arg \ \min_{H \in \mathcal{H}} I(H|E,B)$$

Princip najkrajšega opisa (MDL)

Po Bayesovem teoremu velja:

$$I(H|B,E) = I(E|H,B) + I(H|B) - I(E|B)$$

$I(E|B)$ je konstanta, neodvisna od hipoteze:

$$H_{opt} = arg \ \min_{H \in \mathcal{H}} (I(E|H,B) + I(H|B))$$

Optimalna hipoteza je

- Točna – majhna napaka I(E|H,B)
- Preprosta – majhen I(H|B)

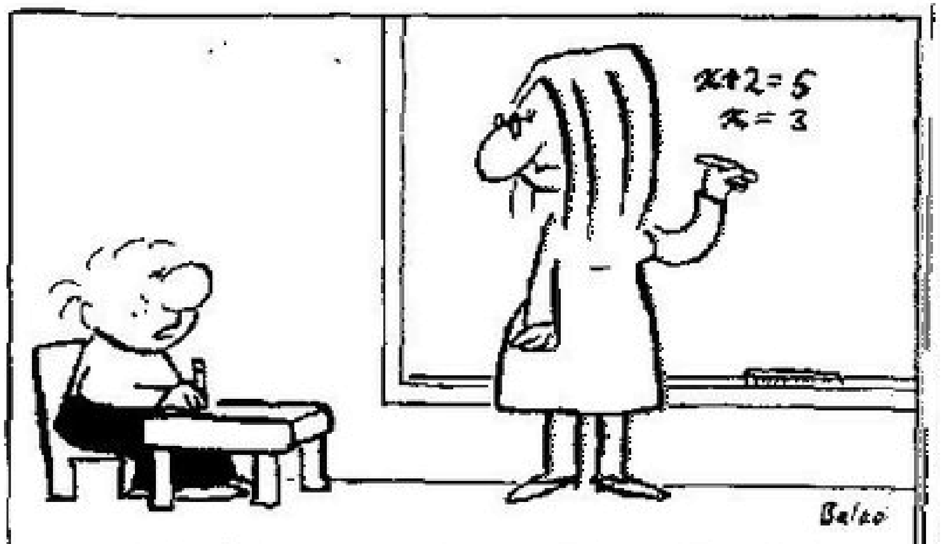
Inkrementalno učenje

Nature never says whether the guesses are correct. Scientific success consists of eventually offering a correct guess and never deviating from it thereafter.

(Osherson et al., 1986)

- Spreminjanje teorije po vsakem novem učnem primeru
- Problem: najmanjša potrebna sprememba trenutne teorije
- Zapominjanje/pozabljanje učnih primerov
- Učni algoritem nikoli z gotovostjo ne ve, če se je naučil optimalno teorijo.

Incremental Learning



"Just a darn minute! Yesterday you said X equals two!"

An Example From Sports Betting

Odds offered by a known online bookmaker for a UEFA Championship League quarter-final match between Inter and Manchester (home, draw, and away):

Inter Milan	2.30	3.10	2.90	Manchester United

Odds tell us what the payout for an individual outcome is. For example, betting 1€ on Inter will pay 2.3€. If they win, of course.

Odds also imply what the teams' chances of winning are. The odds above suggest that Inter is a slight favourite with a 43% chance of winning (1 / 2.30). Manchester, on the other hand, has a 34% chance.

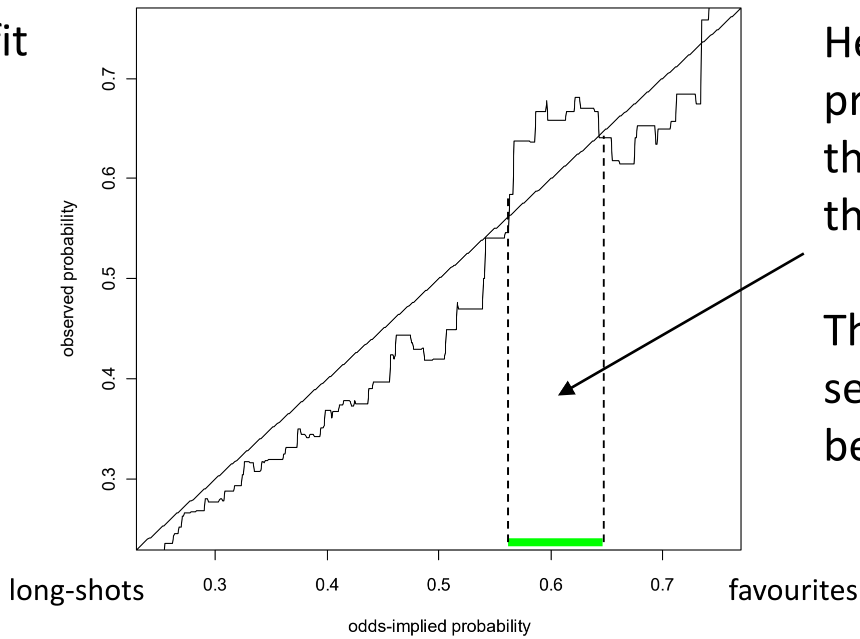
Task:
- A bookmaker offers us betting odds for soccer matches
- Bookmakers can make mistakes and publish favourable odds
- Can we beat the bookmaker and make money by betting on a particular odds band?

We Can Learn from Past Examples

A nearest neighbor approach is used to estimate the probability of the home team winning:

Using the odds and outcomes of 1000 past matches, we get:

The diagonal line indicates zero-profit opportunities.



Here the observed probability is larger than the probability implied by the odds.

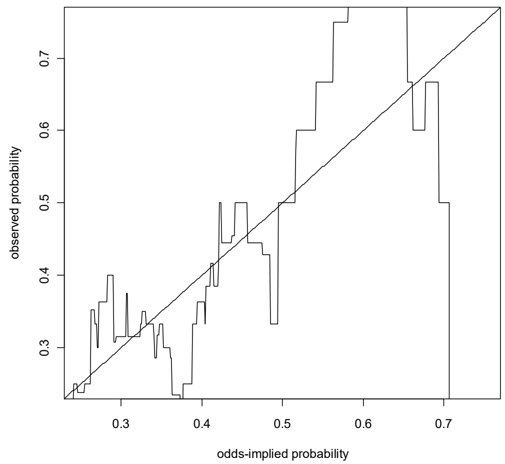
Therefore, this odds band seems to be a profitable betting opportunity.

Bet on the home team whenever the offered odds are between 1.5 and 1.8 !!!

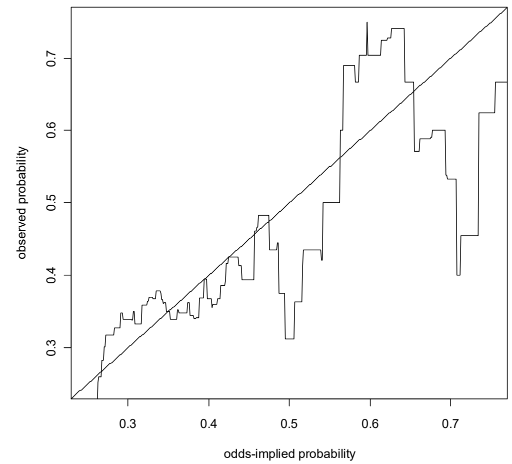
Our Estimation Improves over Time

However,...

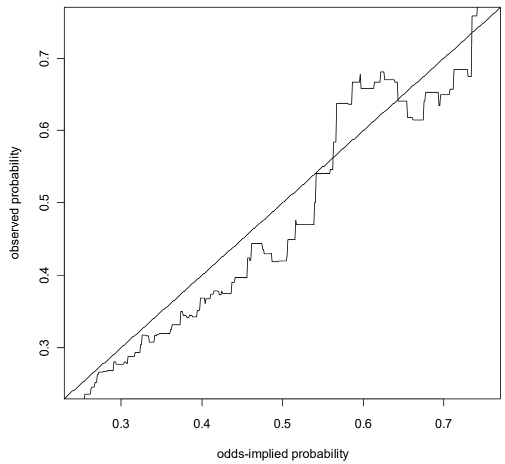
100 matches



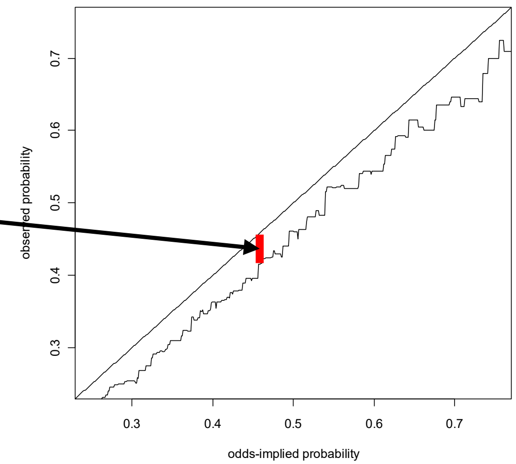
500 matches



1000 matches

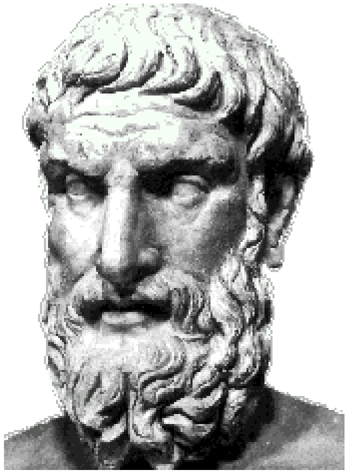


6000 matches



The bookmaker's profit margin.

Princip večkratne razlage



If more than one theory is consistent with the observations, keep all theories. (Epicurus)

For each particular model find what its prediction is and then weight its prediction with the probability of that model, and the weighted sum of those predictions is the optimal prediction. (Peter Cheeseman)

- Obdrži vse konsistentne (verjetne) hipoteze z vhodnimi podatki!

Princip večkratne razlage

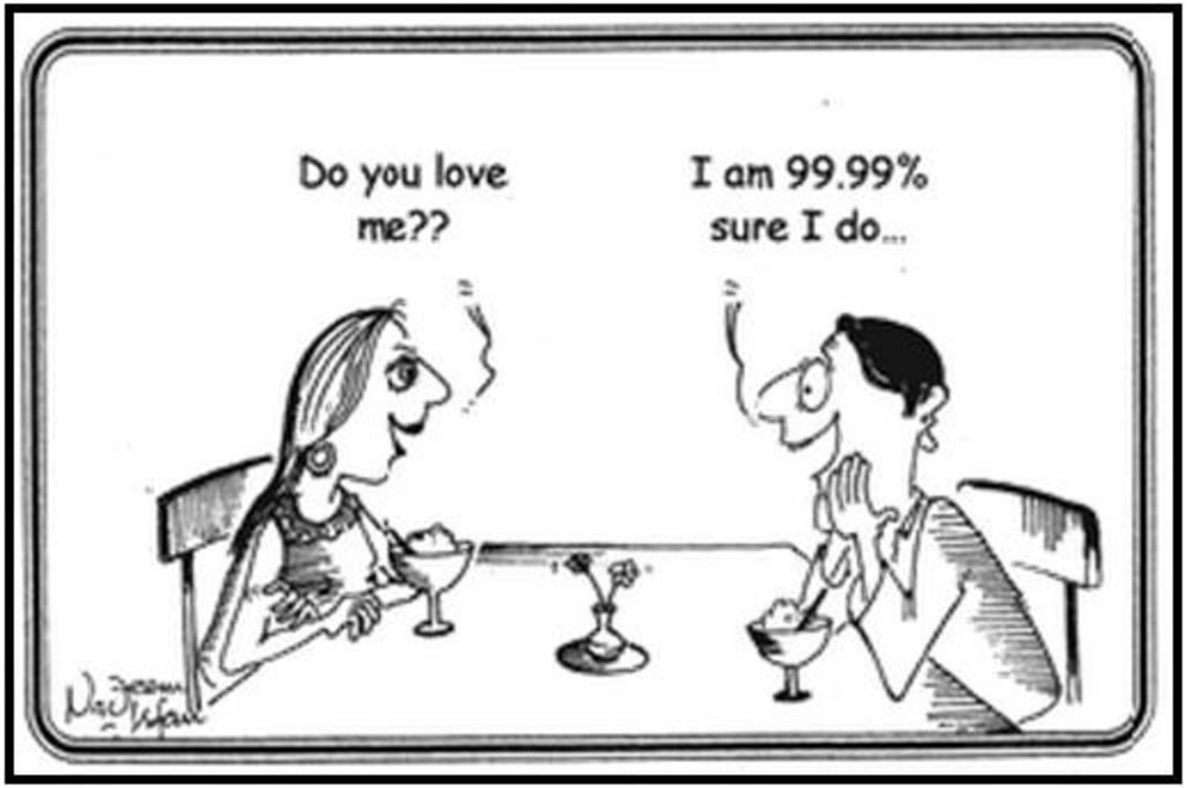
Princip MDL:

- iskanje (v povprečju) najboljše hipoteze
- usmerjanje učnega algoritma

Princip večkratne razlage:

- kombinacija več verjetnih hipotez
- usmerjanje izvajalnega algoritma

Ocenjevanje verjetnosti



Ocenjevanje verjetnosti

oceniti verjetnost iz majhne množice vhodnih podatkov

Apriorna gostota verjetnosti: beta porazdelitev $\beta(a, b)$:

$$p(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & 0 \leq x \leq 1 \\ 0 & sicer \end{cases}$$

$$B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$$

$a > 0$ in $b > 0$ interpretiramo: a uspešnih in b neuspešnih

matematično upanje slučajne spremenljivke p porazdeljene po $\beta(a, b)$:

$$Exp(p) = \frac{a}{a+b}$$

Ocenjevanje verjetnosti

r uspešnih in n vseh primerov, potem ocenimo verjetnost uspeha:

$$p = \frac{r+a}{n+a+b} \qquad \longleftarrow \beta(a+r, b+n-r)$$

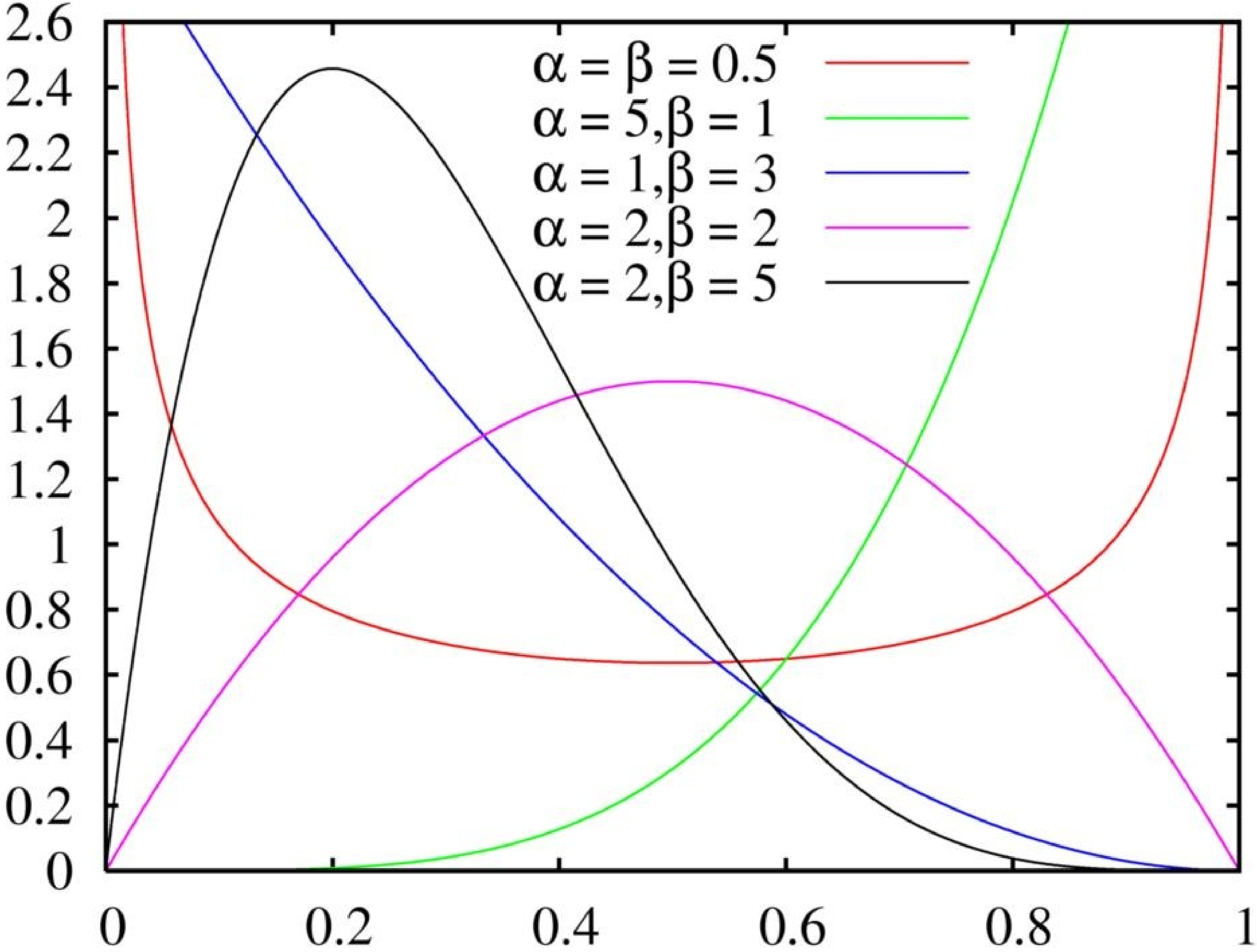
- začetna porazdelitev $\beta(0,0)$: *relativna frekvenca*

$$p = \frac{r}{n}$$

- začetna porazdelitev $\beta(1,1)$: *Laplaceov zakon zaporednosti*
  k = število možnih izidov:

$$0 < p = \frac{r+1}{n+k} < 1$$

Ocenjevanje verjetnosti



Ocenjevanje verjetnosti

- $m = a + b$, $p_0 = \frac{a}{a+b}$: *m-ocena verjetnosti*

$$p = \frac{r + mp_0}{n + m} = \frac{n}{n + m} \times \frac{r}{n} + \frac{m}{n + m} \times p_0$$

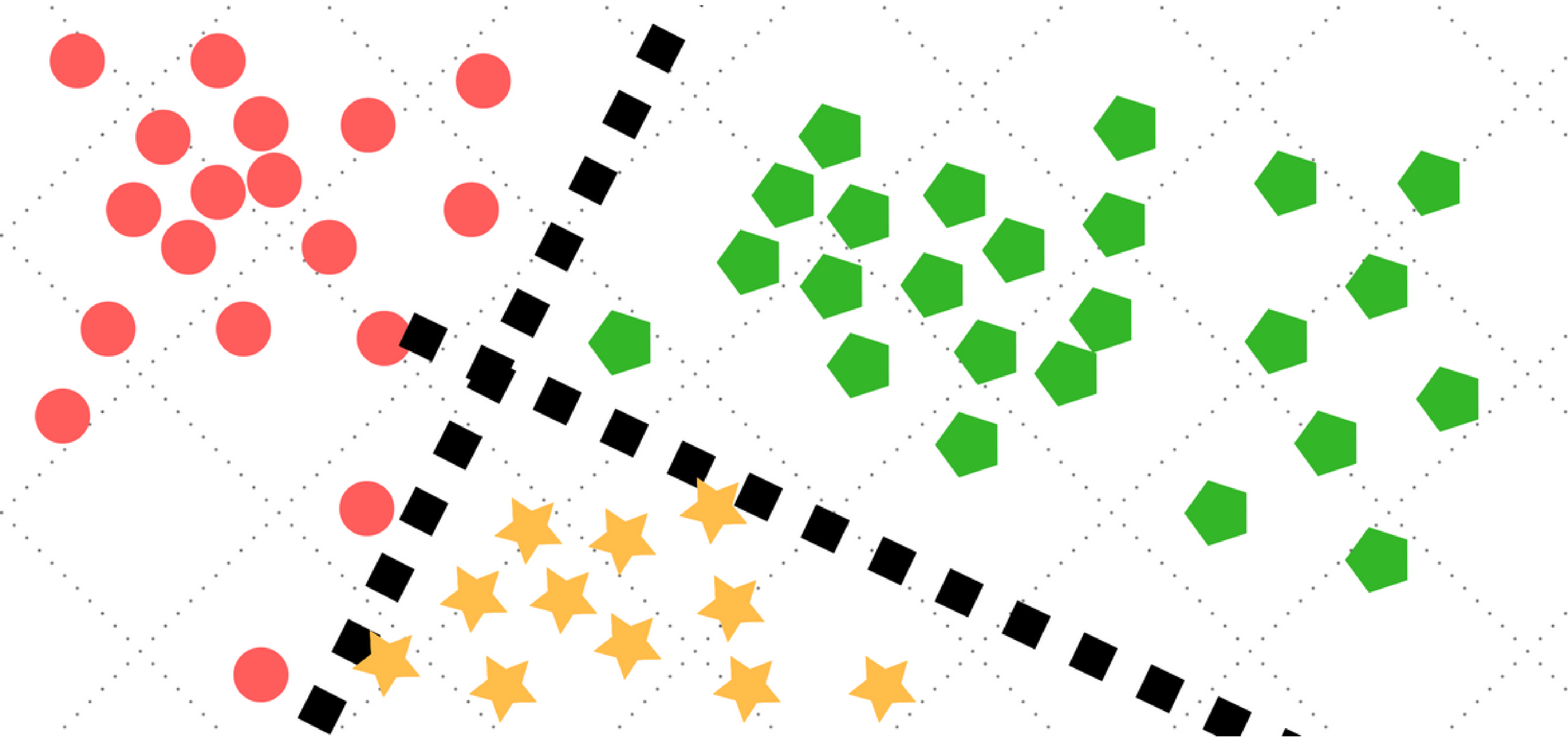
$$m = k,\ p_0 = 1/k \longrightarrow \text{Laplace}$$

Ocenjevanje učenja

Klasifikacija:
- klasifikacijska točnost,
- tabela napačnih klasifikacij,
- cena napačne klasifikacije,
- Brierjeva mera,
- informacijska vsebina,
- senzitivnost in specifičnost, krivulja ROC,

Regresija:
- Srednja kvadratna napaka
- Relativna srednja kvadratna napaka
- Srednja absolutna napaka
- Relativna srednja absolutna napaka

Klasifikacija



Klasifikacijska točnost

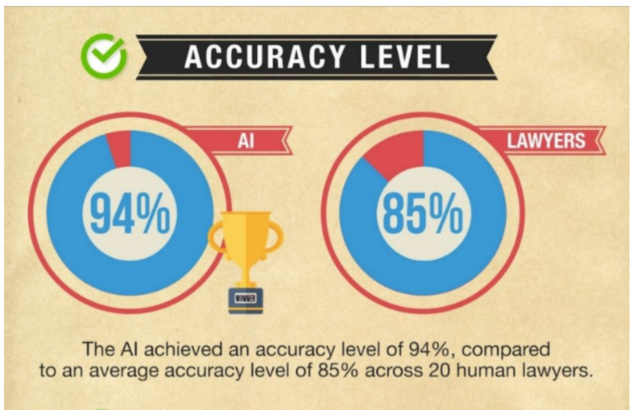
M_R - število možnih razredov

N - število vseh možnih primerov problemov na danem področju

$N^{(p)}$ - število pravilnih rešitev primerov

N_t - število vseh testnih primerov

$$T = \frac{N^{(p)}}{N} \times 100\% \sim T_t = \frac{N_t^{(p)}}{N_t} \times 100\%$$



Klasifikacijska točnost

Klasifikacijska točnost na N_u učnih primerih = zgornja meja:

$$T_u = \frac{N_u^{(p)}}{N_u} \times 100\%$$

Če $T_u >> T_t$:*preveč prilagojeno* učni množici (overfitting)

Klasifikacijska točnost

večinski razred: spodnja meja klasifikacijske točnosti

$N_u^{(i)}$ - število učnih primerov iz i-tega razreda

$$T_v = \max_i \frac{N_u^{(i)}}{N_u}$$

$T_t < T_v$: klasifikator je neuporaben

Problem prognostike ponovitve raka na dojki: $T_v = 80\%$

Večina parametrov je nepomembnih za prognozo.

Apriorni verjetnosti razredov: $P_1 = 0.8$ in $P_2 = 0.2$

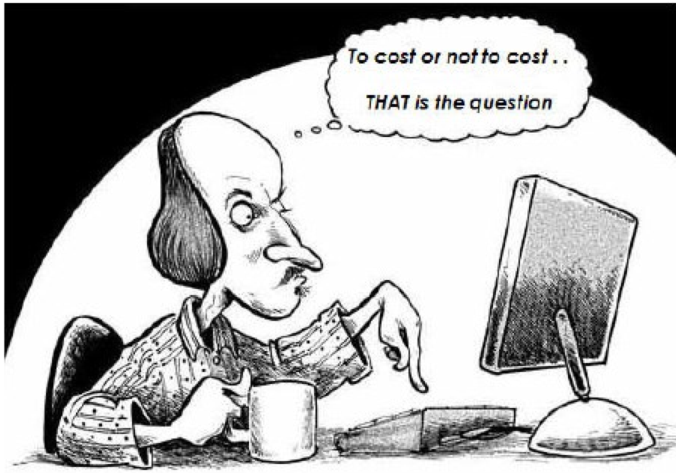
Klasifikacijska točnost: $P_1^2 + P_2^2 = 68\%$

Tabela napak/zmot

pravi razred	klasificiran kot C1	C2	C3	vsota
C1	12.3	2.4	8.5	23.2
C2	5.5	58.7	2.1	66.3
C3	0.0	2.0	8.5	10.5
vsota	17.8	63.1	19.1	100.0

Povprečna cena napak

$N_t^{(ij)}$ - število testnih primerov iz i-tega razreda, ki jih dana teorija klasificira v j-ti razred.



$$N_t^{(p)} = \sum_{i=1}^{M_R} N_t^{(ii)}$$

C_{ij} - cena napačne klasifikacije

Lahko: $C_{ij} \neq C_{ji}$ Ponavadi: $C_{ii} = 0,\ i = 1, ..., M_R$

Povprečna cena napačne klasifikacije:

$$C_t = \frac{\Sigma_{i,j}(C_{ij} N_t^{(ij)})}{N_t}$$

Povprečna cena napak

napake	->C1	->C2	->C3
C1	12,3	2,4	8,5
C2	5,5	58,7	2,1
C3	0,0	2,0	8,5

cene	->C1	->C2	->C3
C1	0	10	1
C2	1	0	2
C3	1	1	0

Cena = 2,4X10 + 8,5X1 + 5,5X1 + 2,1X2 + 0,0X1 + 2,0X1 = 44,2

Brierjeva mera

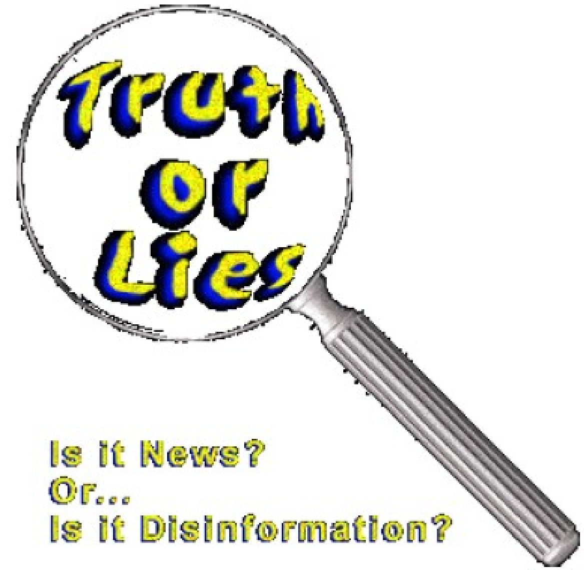
- upošteva napovedane verjetnosti razredov
- povprečna kvadratna napaka napovedanih verjetnosti
- min = 0 (najboljše), max = 2 (najslabše)

M_R število razredov
$r^{(j)}$ razred j-tega testnega primera
$P'_j(r_i), i = 1..M_R$ napovedana verjetnostna distribucija
$C_j(r^{(j)}) = 1$ in $C_j(r_i) = 0, r_i \neq r^{(j)}$ Ciljna distribucija
N_t število testnih primerov:

$$Brier = MSE_P = \frac{\sum_{j=1}^{N_t} \sum_{i=1}^{M_R} (C_j(r_i) - P'_j(r_i))^2}{N_t}$$

kvaliteta klasifikatorja: $1 - Brier/2$.

Informacijska vsebina odgovora

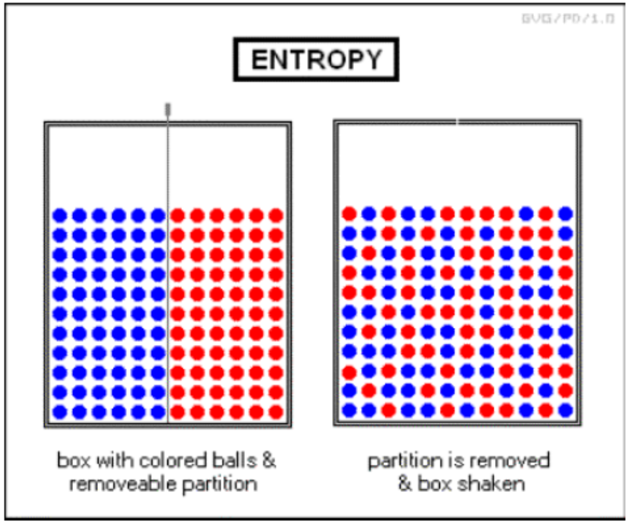


Odgovor: verjetnostna distribucija po vseh razredih

Apriorne verjetnosti in aposteriorne verjetnosti razredov

domena	$T_{t\iota}$	T_v	M_r	$H(R)$	Inf
rak na dojki	80%	80%	2	0.72 bit	0.0 bit
primarni tumor	45%	25%	22	3.64 bit	1.6 bit

Entropija



Apriorna verjetnost i-tega razreda: $P(r_i) = \frac{N_u^{(i)}}{N_u}$

Informacija, da zvemo, da primer pripada i-temu razredu:

$$H(r_i) = -\log_2 P(r_i) \quad [bit]$$

Entropija razredov: $H(R) = -\sum_{i=1}^{R} (P(r_i) \log_2 P(r_i))$

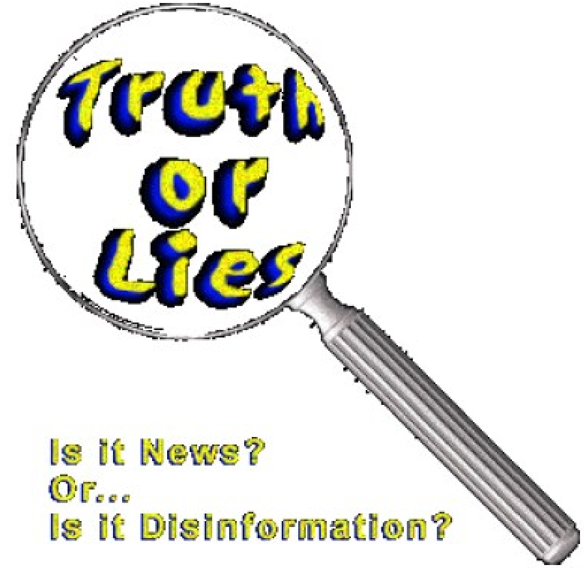
$H(R)$ doseže maksimum, ko so vsi razredi enako verjetni:

$$P(r_i) = \frac{1}{M_R}, \qquad i = 1, ..., M_R$$

ter minimum, ko so vsi primeri iz istega razreda r_j:

$$P(r_j) = 1, \qquad in \qquad P(r_i) = 0, \quad i \neq j$$

Informacijska vsebina odgovora



Povprečna informacijska vsebina odgovora (information score):

$r^{(j)}$ - pravilni razred danega j-tega testnega primera

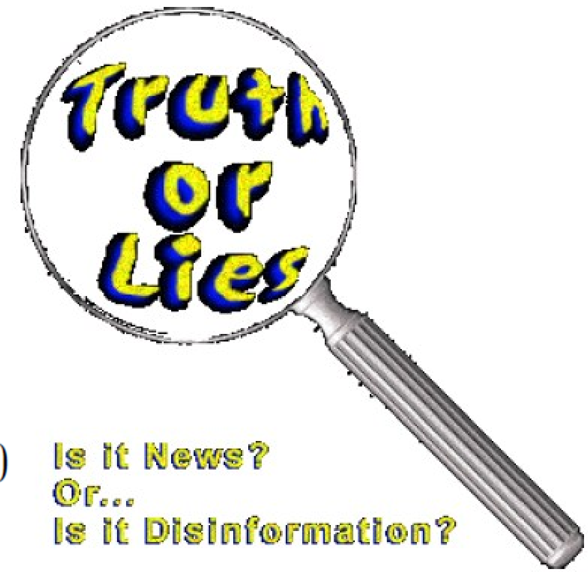
$P'(r^{(j)})$ - aposteriorna verjetnost tega razreda

$$Inf = \frac{\Sigma_{j=1}^{N_t} Inf_j}{N_t} \qquad [bit]$$

in

$$Inf_j = \begin{cases} -\log_2 P(r^{(j)}) + \log_2 P'(r^{(j)}), & P'(r^{(j)}) \geq P(r^{(j)}) \\ -(-\log_2(1 - P(r^{(j)})) + \log_2(1 - P'(r^{(j)}))), & P'(r^{(j)}) < P(r^{(j)}) \end{cases}$$

Informacijska vsebina: klasifikacijska točnost



Imejmo dva možna razreda: $P(r_1) = 0.8$ in $P(r_2) = 0$
Klasifikator vrne $P'(r_1) = 0.6$ in $P'(r_2) = 0.4$

- Če je pravilni razred r_1:
  klasifikacijska točnost: odgovor je pravilen
  informacijska vsebina: odgovor je nepravilen (zavajajoč)

- Če je pravilni razred r_2:
  klasifikacijska točnost: odgovor je nepravilen
  informacijska vsebina: odgovor je pravilen (koristen)

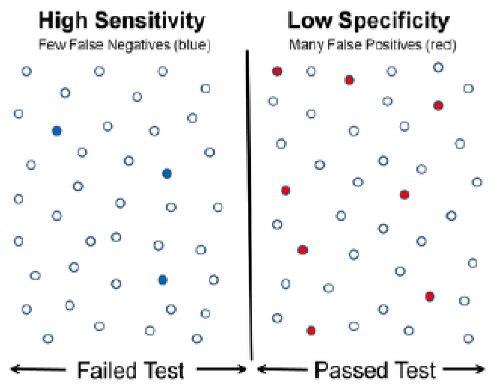
Relativna informacijska vsebina

$$\text{Meje: } 0 \leq Inf \leq H(R)$$

Relativna informacijska vsebina odgovora:

$$RInf = \frac{Inf}{H(R)} \times 100\%$$

Senzitivnost in specifičnost



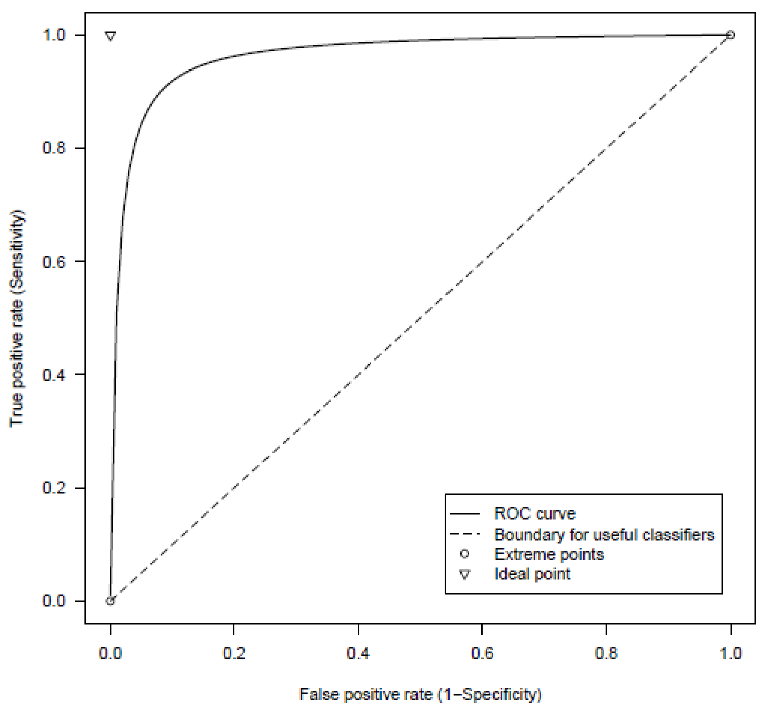
pravi razred	klasificiran kot P	N	vsota
P	TP	FN	POS=TP+FN
N	FP	TN	NEG=FP+TN
vsota	PP=TP+FP	PN=FN+TN	n = TP+FP+FN+TN

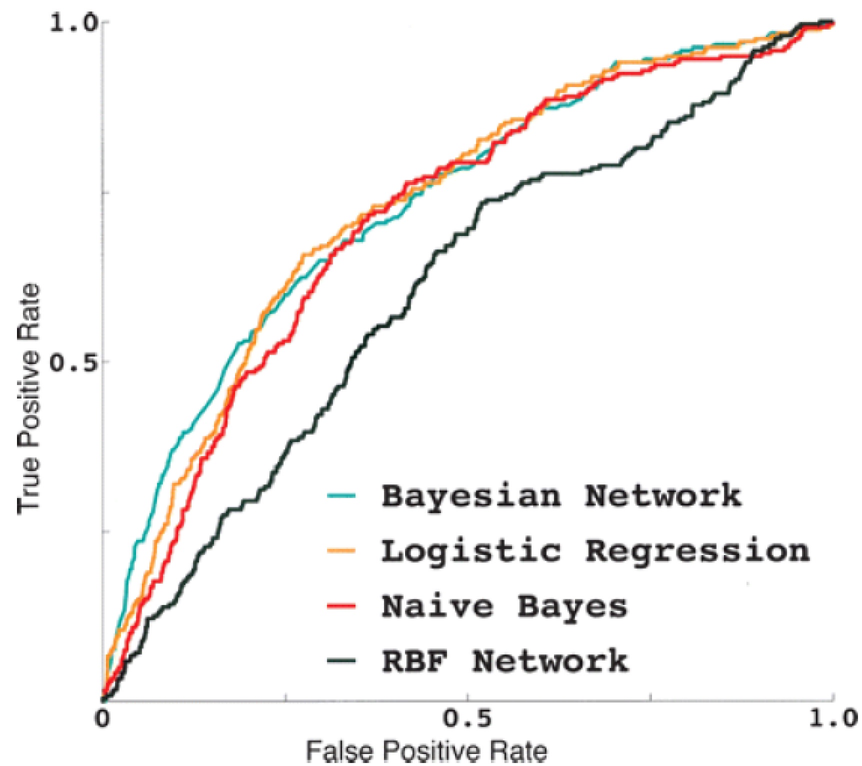
$$Senzitivnost = \frac{TP}{TP+FN} = \frac{TP}{POS}$$

$$Specificnost = \frac{TN}{TN+FP} = \frac{TN}{NEG}$$

$$Tocnost = \frac{TP+TN}{TN+FP+FN+TN} = \frac{TP+TN}{n}$$

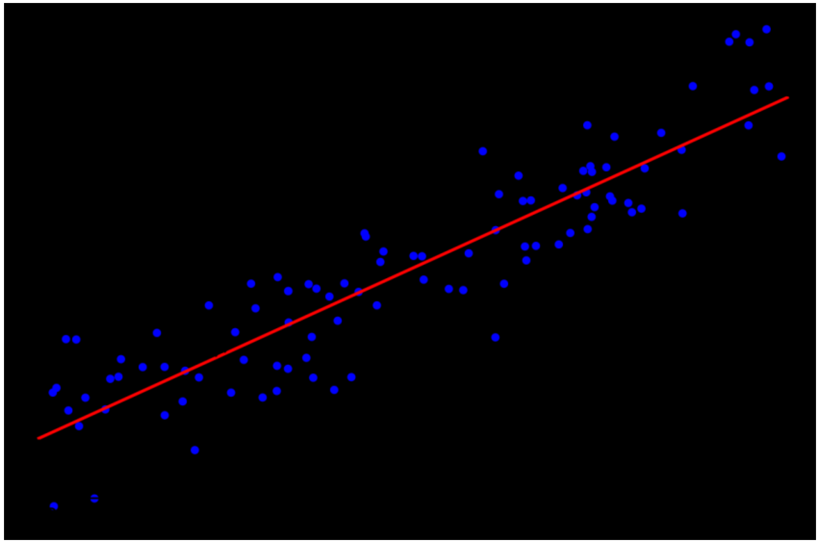
Krivulja ROC
(Reciever Operating Characteristic curve)

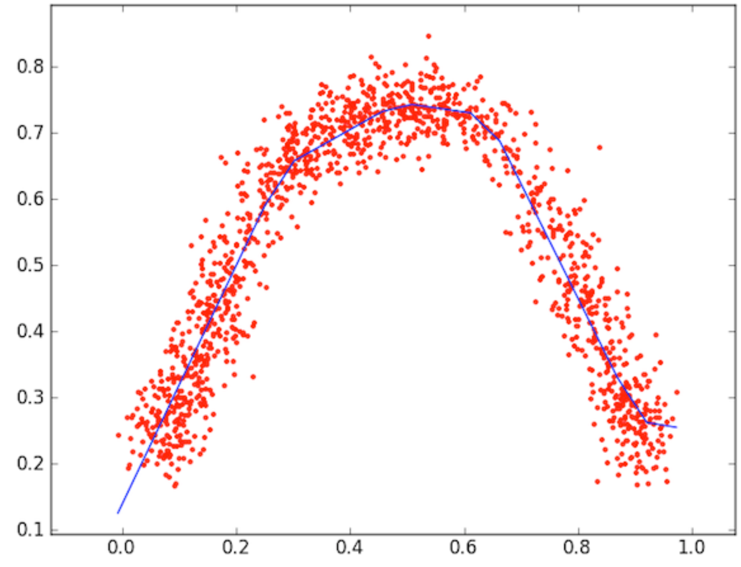


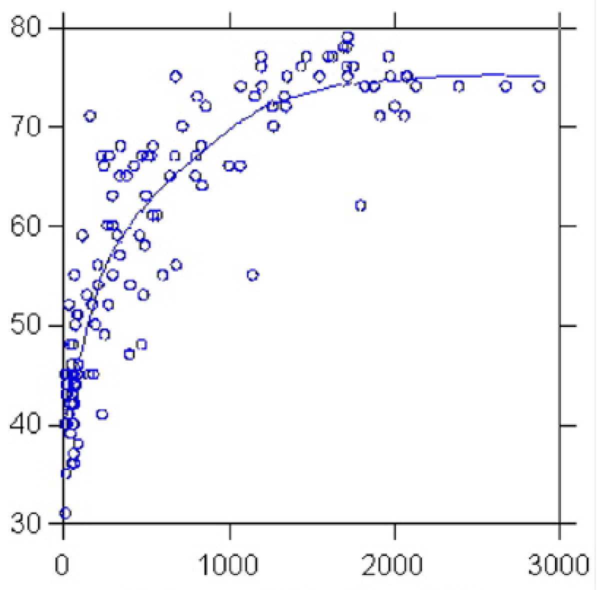


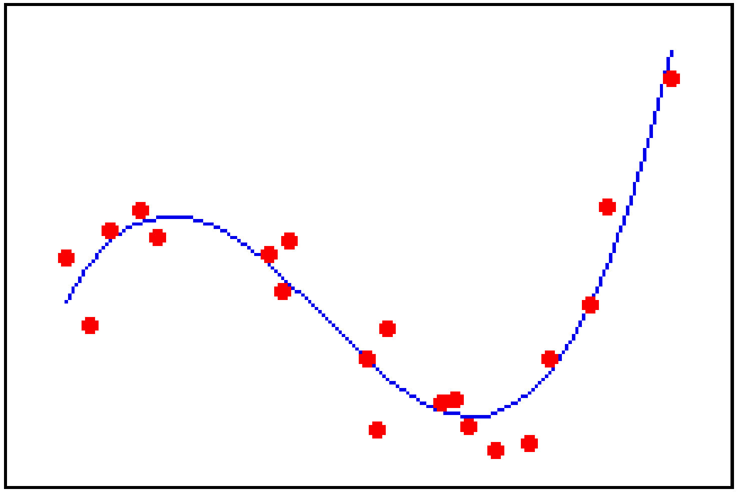
ploščina pod krivuljo ROC (Area Under the ROC Curve, AUC)

Regresija

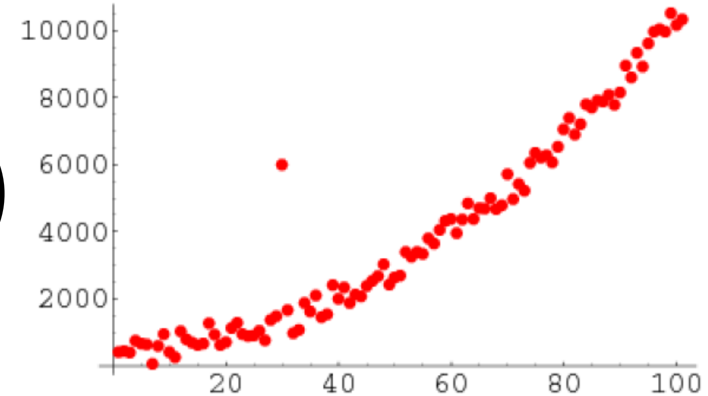



0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.2
0.4
0.6
0.8
1.0


80
70
60
50
40
30
0
1000
2000
3000



Srednja kvadratna napaka (mean squared error, MSE) Relativna MSE



Pri regresijskih problemih:

napovedana vrednost: $\hat{f}(i)$ želena vrednost: $f(i)$

$$E = \frac{1}{N}\sum_{i=1}^{N}(f(i) - \hat{f}(i))^2$$

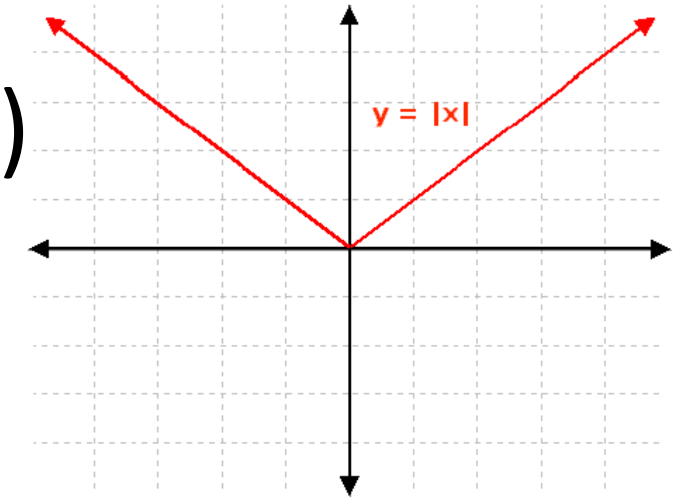
Relativna srednja kvadratna napaka:

$$0 \leq RE = \frac{N \times E}{\sum_i (f(i) - \overline{f})^2} \leq 1 \quad kjer\ je\ \ \overline{f} = \frac{1}{N}\sum_i f(i)$$

$\hat{f}(i) = \overline{f} \longrightarrow RE = 1$ trivialna f. $(RE > 1 \rightarrow$ neuporabna)

$\hat{f}(i) = f(i) \longrightarrow RE = 0$ idealna funkcija

Srednja absolutna napaka (mean absolute error, MAE) **Relativna MAE**



mean absolute error (MAE)

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |f(i) - \hat{f}(i)|$$

relativna srednja absolutna napaka:

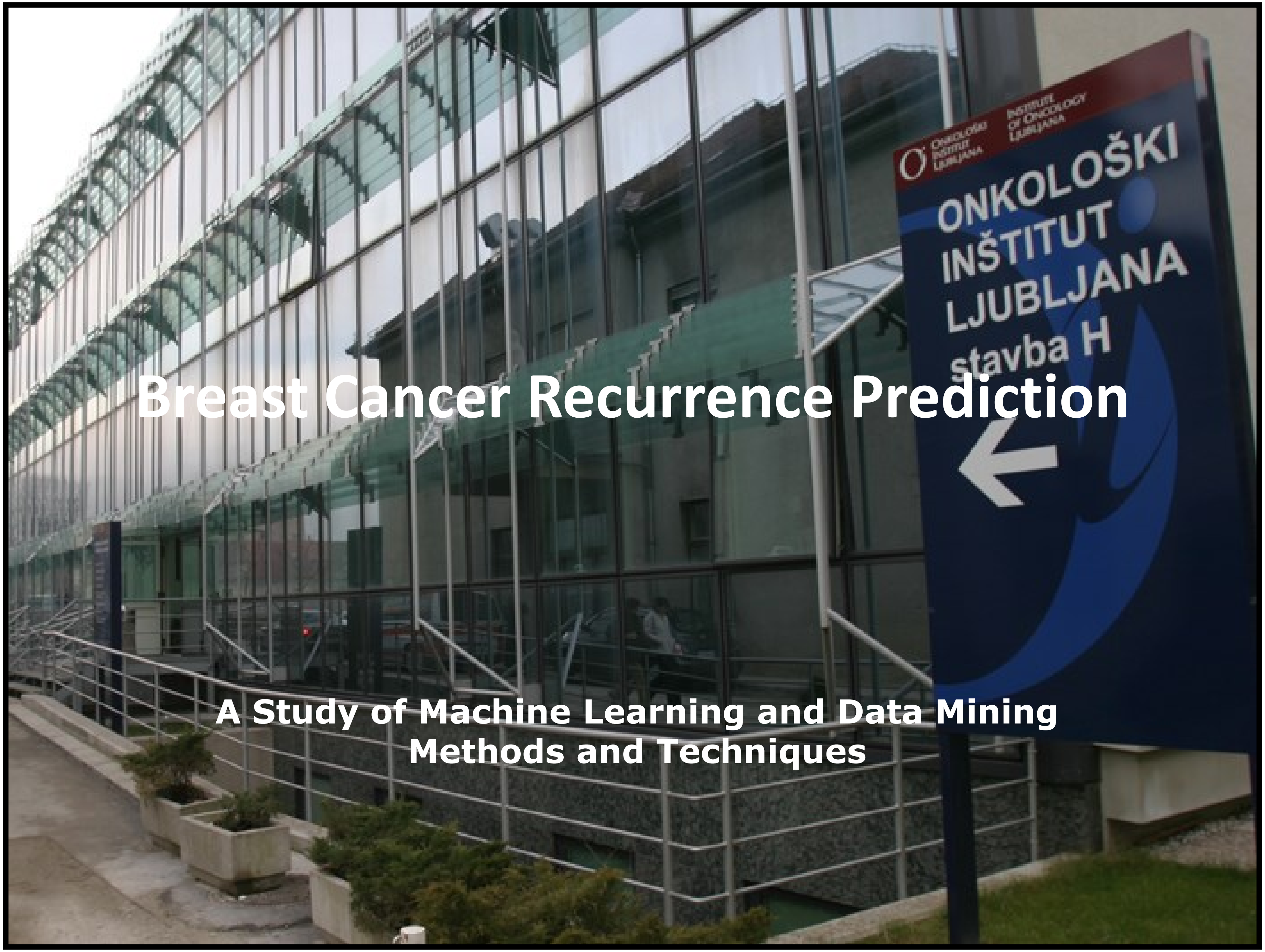
$$RMAE = \frac{N \times MAE}{\sum_i |f(i) - \overline{f}|}$$

$$0 \leq RMAE \leq 1$$

Machine Learning...

"It guessed your PIN number and cleaned out your bank account. Your move."

Breast Cancer Recurrence Prediction



A Study of Machine Learning and Data Mining Methods and Techniques

The Data

Provided by the Institute of Oncology, Ljubljana

Post-surgery data for about 1000 breast cancer patients.

+

Recurrence and time of recurrence.

The Data

	class1	class2	menop	stage	grade	hType	PgR	inv	nLymph	cTh	hTh	famHist	LVI	ER	maxNode	posRatio	age
300	11.82	0	1	2	2	1	0	0	1	1	0	3	0	1	2	3	2
301	4.89	1	0	1	2	1	0	0	2	1	0	0	0	2	1	4	3
302	14.63	0	1	1	4	2	0	0	0	0	0	1	0	1	1	1	3
303	21.83	0	0	1	4	2	1	0	1	0	0	9	0	4	1	2	2
304	19.87	0	0	1	2	1	0	0	0	0	0	0	0	1	2	1	2
305	7.54	0	1	2	3	1	9	2	1	0	1	1	0	3	3	3	4
306	15.15	0	0	1	4	2	1	0	0	0	0	2	0	4	1	1	2
307	0.30	1	0	2	2	1	0	0	3	0	0	9	0	1	1	4	2
308	12.49	0	1	2	2	3	1	0	0	0	0	0	0	4	1	1	5
309	1.77	1	0	2	3	1	1	2	2	1	0	9	1	3	3	3	2

Each patient is described with 17 values:
- 15 patient's features
- 2 values, which describe the outcome

1 instance = 1 patient

	class1	class2	menop	stage	grade	hType	PgR	inv	nLymph	cTh	hTh	famHist	LVI	ER	maxNode	posRatio	age
300	11.82	0	1	2	2	1	0	0	1	1	0	3	0	1	2	3	2
301	4.89	1	0	1	2	1	0	0	2	1	0	0	0	2	1	4	3
302	14.63	0	1	1	4	2	0	0	0	0	0	1	0	1	1	1	3
303	21.83	0	0	1	4	2	1	0	1	0	0	9	0	4	1	2	2
304	19.87	0	0	1	2	1	0	0	0	0	0	0	0	1	2	1	2
305	7.54	0	1	2	3	1	9	2	1	0	1	1	0	3	3	3	4
306	15.15	0	0	1	4	2	1	0	0	0	0	2	0	4	1	1	2
307	0.30	1	0	2	2	1	0	0	3	0	0	9	0	1	1	4	2
308	12.49	0	1	2	2	3	1	0	0	0	0	0	0	4	1	1	5
309	1.77	1	0	2	3	1	1	2	2	1	0	9	1	3	3	3	2

- Menopause?
- Tumor stage
- Tumor grade
- Histological type
- Progesterone receptor lvl.
- Invasive tumor type
- Number of positive lymph nodes

- Hormonal therapy?
- Chemotherapy?
- Family medical history
- Lymphovascular invasion?
- Estrogen receptor lvl.
- Size of max. removed node
- Ratio of positive lymph nodes
- Age group

Prognostic Features

	class1	class2	menop	stage	grade	hType	PgR	inv	nLymph	cTh	hTh	famHist	LVI	ER	maxNode	posRatio	age
300	11.82	0	1	2	2	1	0	0	1	1	0	3	0	1	2	3	2
301	4.89	1	0	1	2	1	0	0	2	1	0	0	0	2	1	4	3
302	14.63	0	1	1	4	2	0	0	0	0	0	1	0	1	1	1	3
303	21.83	0	0	1	4	2	1	0	1	0	0	9	0	4	1	2	2
304	19.87	0	0	1	2	1	0	0	0	0	0	0	0	1	2	1	2
305	7.54	0	1	2	3	1	9	2	1	0	1	1	0	3	3	3	4
306	15.15	0	0	1	4	2	1	0	0	0	0	2	0	4	1	1	2
307	0.30	1	0	2	2	1	0	0	3	0	0	9	0	1	1	4	2
308	12.49	0	1	2	2	3	1	0	0	0	0	0	0	4	1	1	5
309	1.77	1	0	2	3	1	1	2	2	1	0	9	1	3	3	3	2

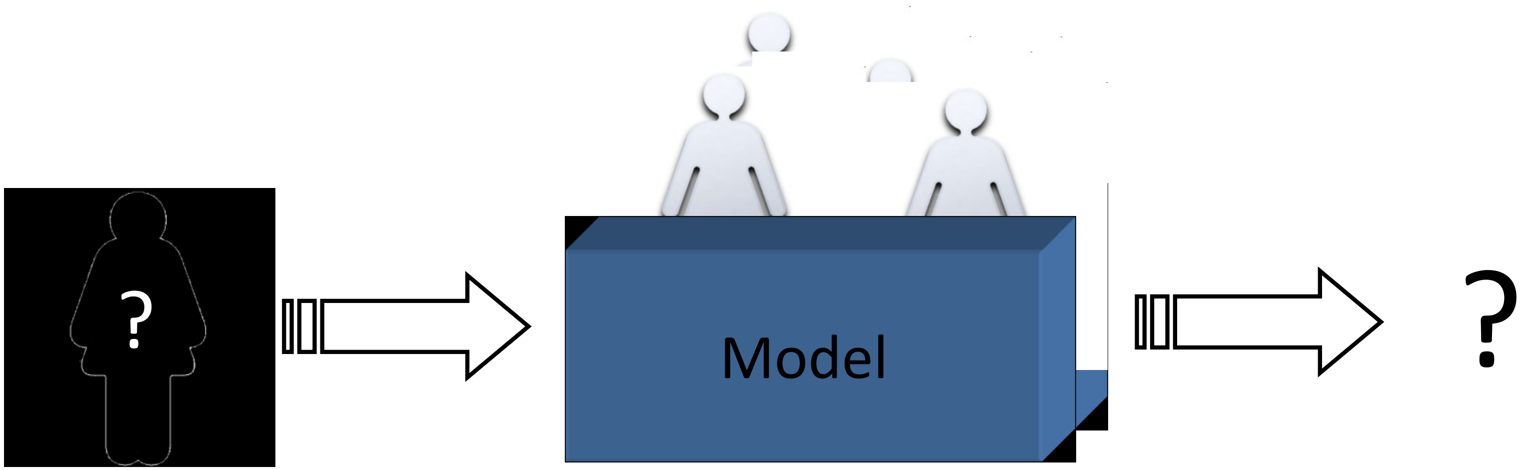
- Menopause?
- Tumor stage
- Tumor grade
- Histological type
- Progesterone receptor lvl.
- Invasive tumor type
- Number of positive lymph nodes

- Hormonal therapy?
- Chemotherapy?
- Family medical history
- Lymphovascular invasion?
- Estrogen receptor lvl.
- Size of max. removed node
- Ratio of positive lymph nodes
- Age group

Oncologists use these attributes for prognosis in every-day medical practice.

Basic Task in ML

We want to learn from past examples,with known outcomes.



To predict the outcome for a new patient.

Let’s Use a Decision Tree

How should we define the outcome?
(we have 2 sub-problems)

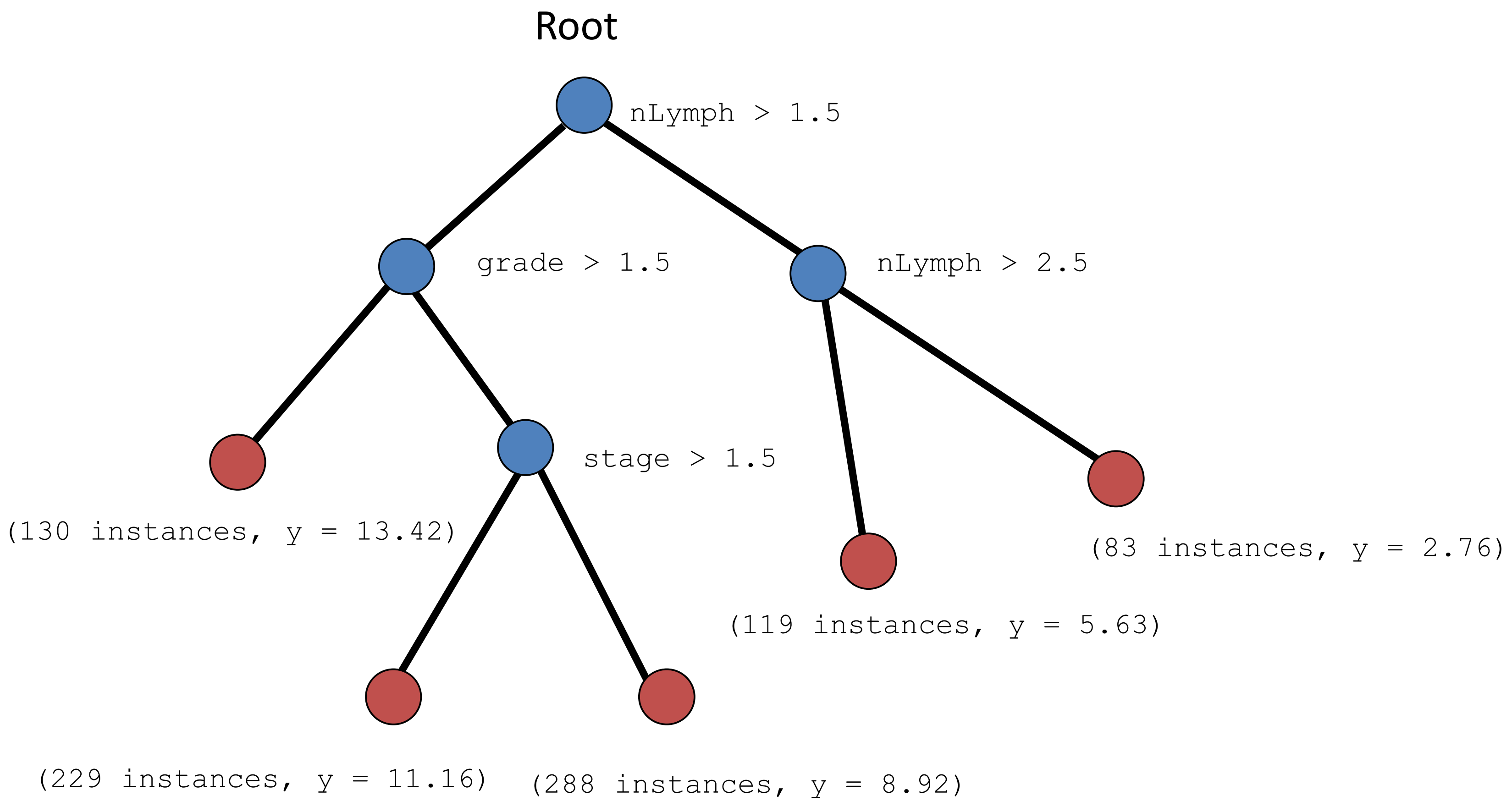
Recurrence?
(yes / no)

CLASSIFICATION TREE

Time to recurrence?
(continuous value)

REGRESSION TREE

A Regression Tree



Basic Evaluation of a Classifier

Our model predicts some instances correctly some incorrectly.

```
...
[40,] 0    0
[41,] 0    0
[42,] 0    0
[43,] 0    0
[44,] 0    0
[45,] 1    1
[46,] 0    0
[47,] 0    0
[48,] 1    0
[49,] 0    0
[50,] 1    0
[51,] 1    1
[52,] 1    0
[53,] 1    1
[54,] 1    1
[55,] 1    1
[56,] 0    1
...
```

$$\text{accuracy} = \frac{\text{number of correctly predicted instances}}{\text{total number of test instances}} = 72\%$$

Accuracy of predicting with the majority class value (0) is approximately 62%.

information score = 0.22 bit

$$\text{relative inf. score} = \frac{\text{information score}}{\text{class entropy}} = 0.23$$

Further Evaluation

Confusion Matrix

		predicted 0	predicted 1
actual value	**0**	57	5
	1	23	15

False Positives

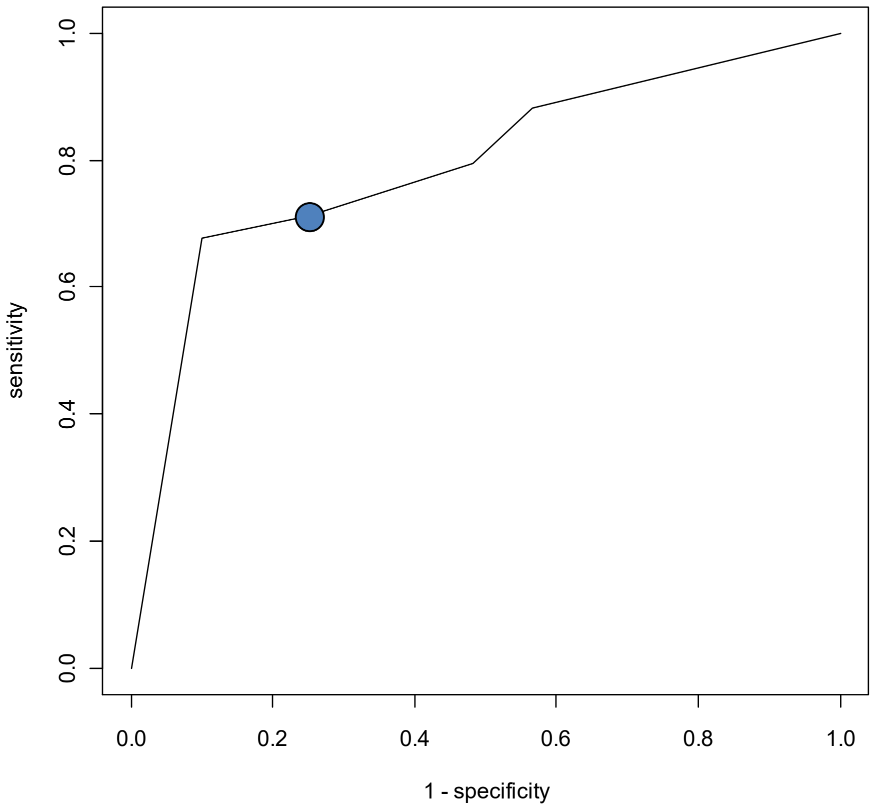
False Negatives

$$\text{sensitivity} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} = 0.75$$

$$\text{specificity} = \frac{\text{True Negatives}}{\text{True Negatives + False Positives}} = 0.712$$

ROC Curve

Receiver Operating Characteristic (ROC) curve.



Evaluating Regression Models

How close are these predicted values (blue) to the actual values (black)?

...

39	16.30116359	8.920783
40	8.95550992	13.420565
41	17.83436003	8.920783
42	17.18275154	13.420565
43	14.54346338	11.157686
44	11.08829569	13.420565
45	1.70841889	5.625156
46	16.58316222	13.420565
47	6.25872690	11.157686
48	4.13689254	13.420565
49	7.34017796	11.157686
50	1.84257358	11.157686
51	1.95208761	5.625156
52	3.50171116	8.920783
53	0.98015058	2.756620

...

mean absolute error (MAE) = 5.01

(on average, the model misses by 5 years)

mean squared error (MSE) = 34.81

Further Evaluation

mean absolute error (MAE) = 5.01

mean squared error (MSE) = 34.81

How good are these results compared to simply predicting with the mean value across training instances?

$$\text{relative mean absolute error (RMAE)} = \frac{\text{MAE}}{\text{MAE of predicting with mean value}} = 0.86$$

MAE of predicting with mean value: $\boxed{5.82}$

$$\text{relative mean squared error (RMSE)} = \frac{\text{MSE}}{\text{MSE of predicting with mean value}} = 0.84$$

MSE of predicting with mean value: $\boxed{41.46}$

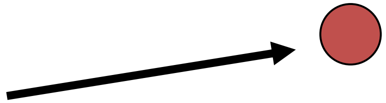
Values lower than 1 indicate that the model is useful.

correlation coefficient = 0.087

Growing a Regression Tree

`Leaves: 1, MSE = 41.46`

Root



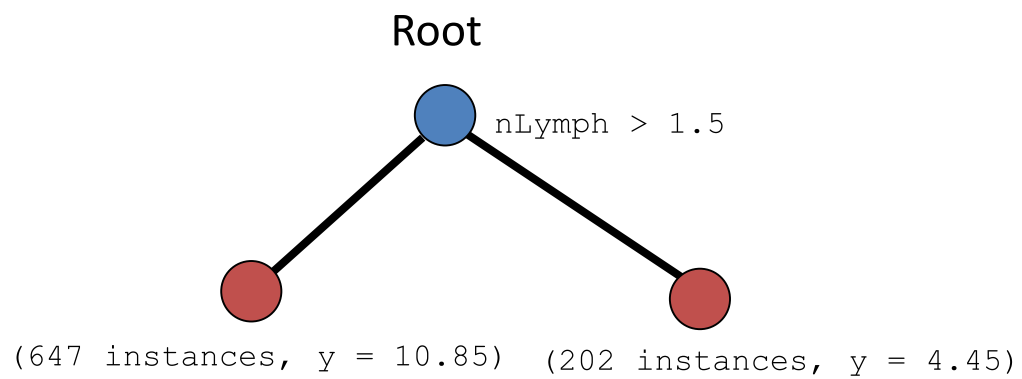
`(849 instances, y = 9.15)`

The most simple tree.

Growing a Regression Tree

Leaves: 1, MSE = 41.46

Leaves: 2, MSE = 36.32

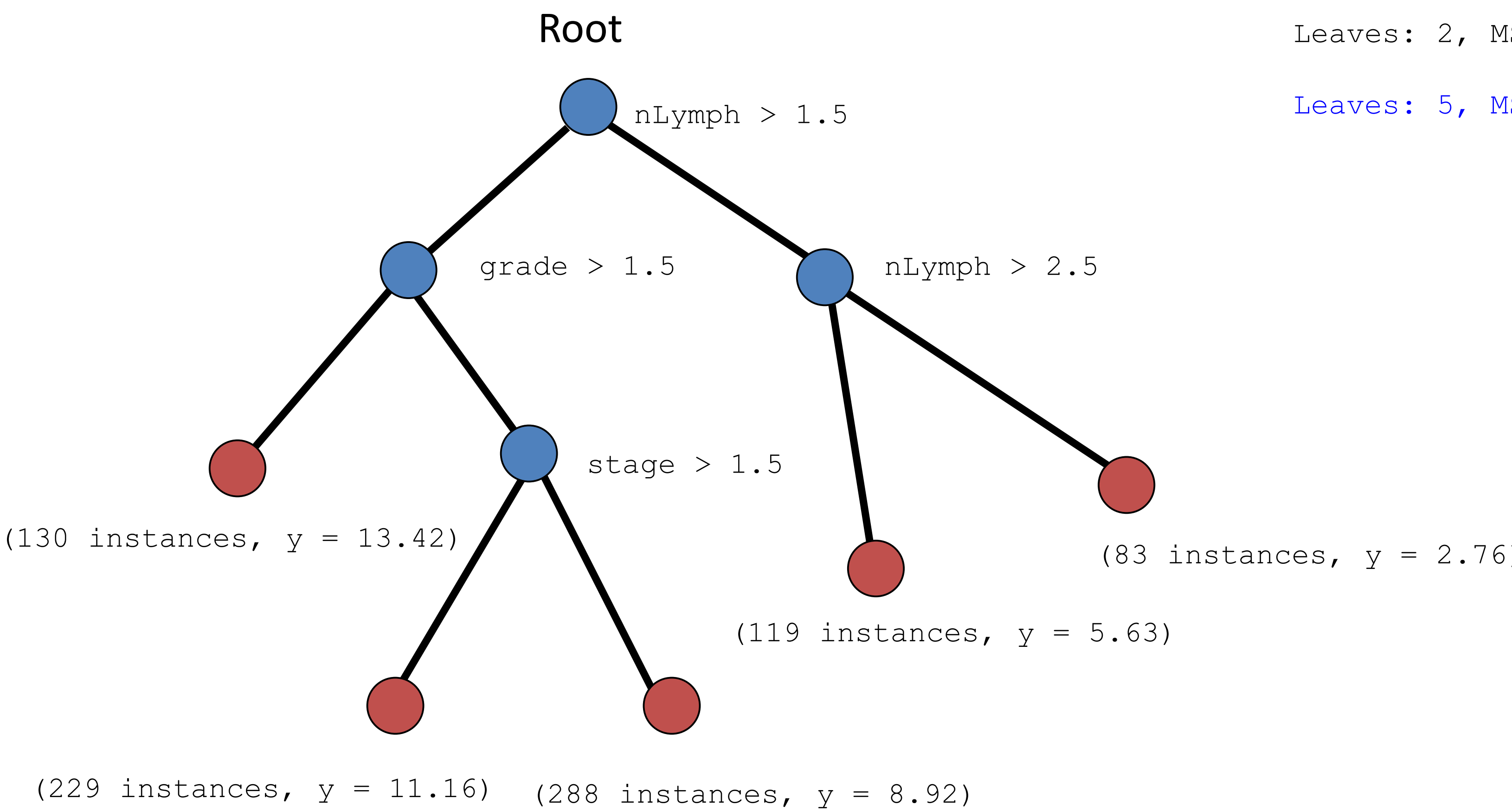


Growing a Regression Tree

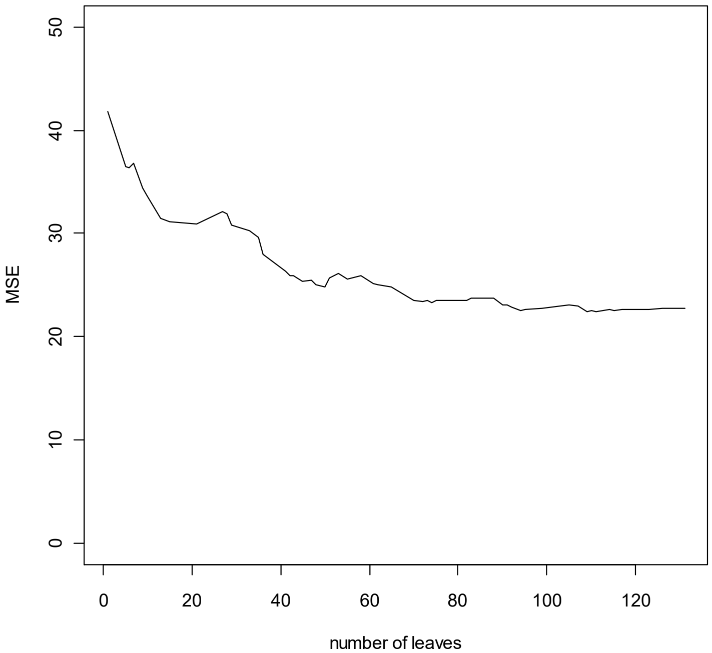
Leaves: 1, MSE = 41.46

Leaves: 2, MSE = 36.32

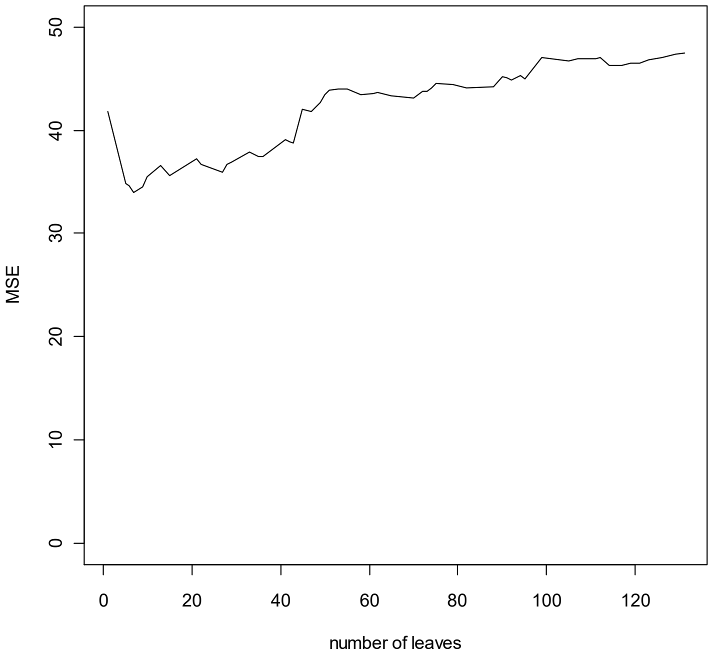
Leaves: 5, MSE = 34.81



Overfitting a Decision Tree



Results on the training data set



Results on the test data set

Further increasing the size of the tree may result in overfitting and a higher error.