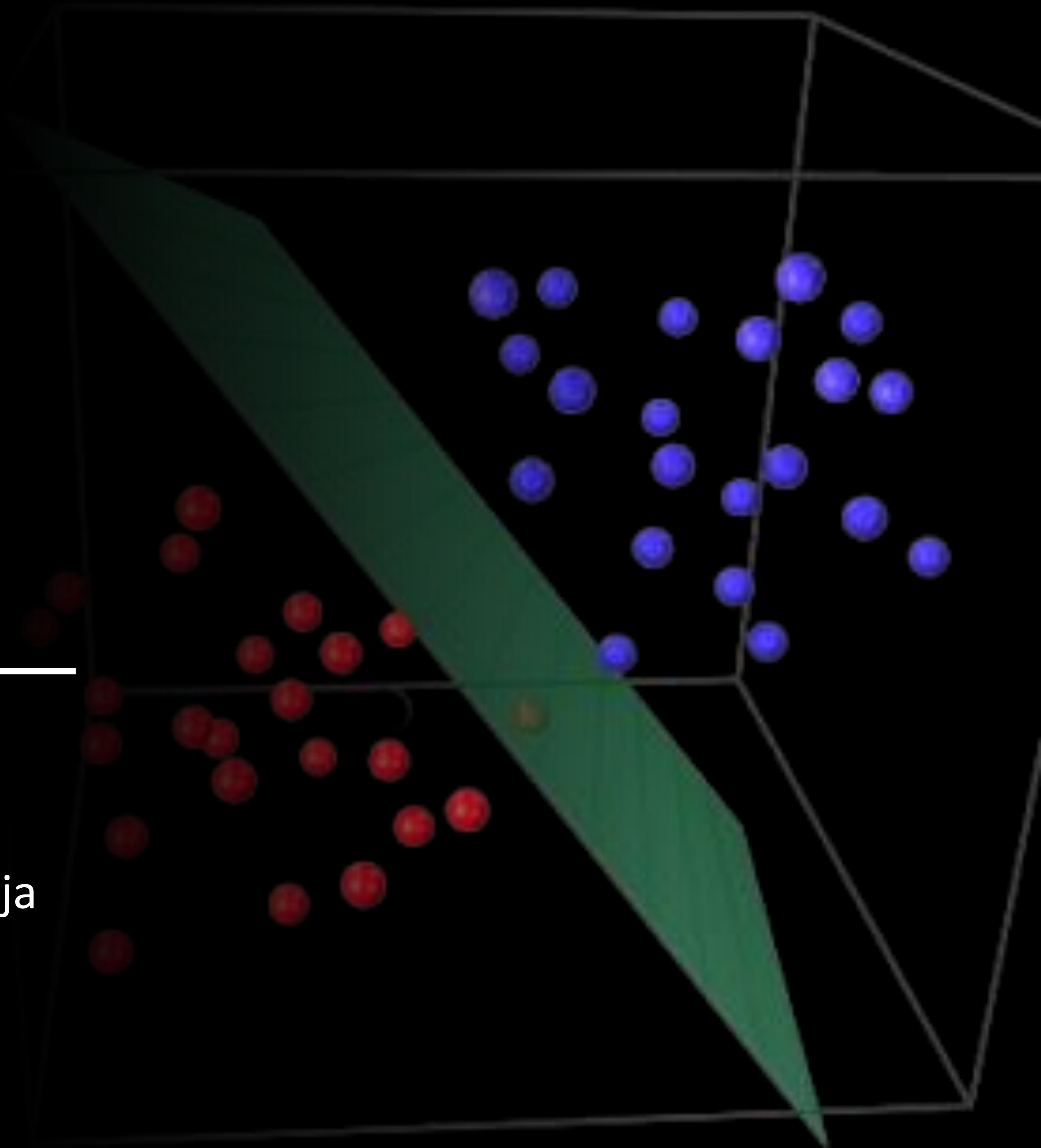
Linearni modeli za klasifikaciju

Logistička regresija

Perceptron

Praktična razmatranja



Klasifikacija

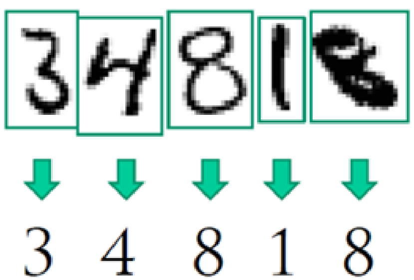
Y uzima kao vrednost jednu od konačno mnogo *diskretnih* kategorija

Binarna:

„Pozitivna klasa“ $y = +1$
„Negativna klasa“ $y = -1$

Višekategorijska:

$y \in \{1,2, \dots , C\}$



Iris setosa *Iris versicolor* *Iris virginica*

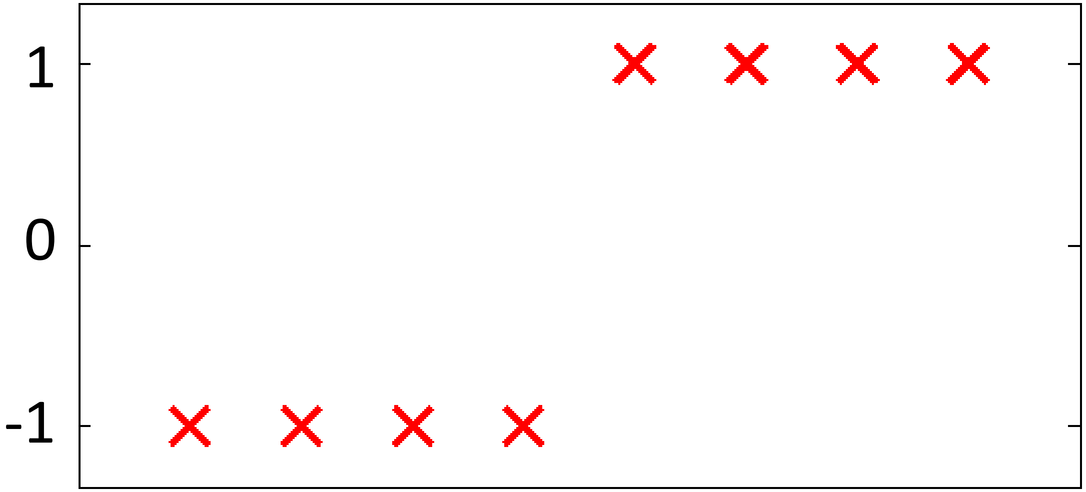
Možemo li primeniti linearnu regresiju?

- Linearna regresija može da „nauči“ funkcije koje imaju realni izlaz $y = f(x) \in \mathbb{R}$

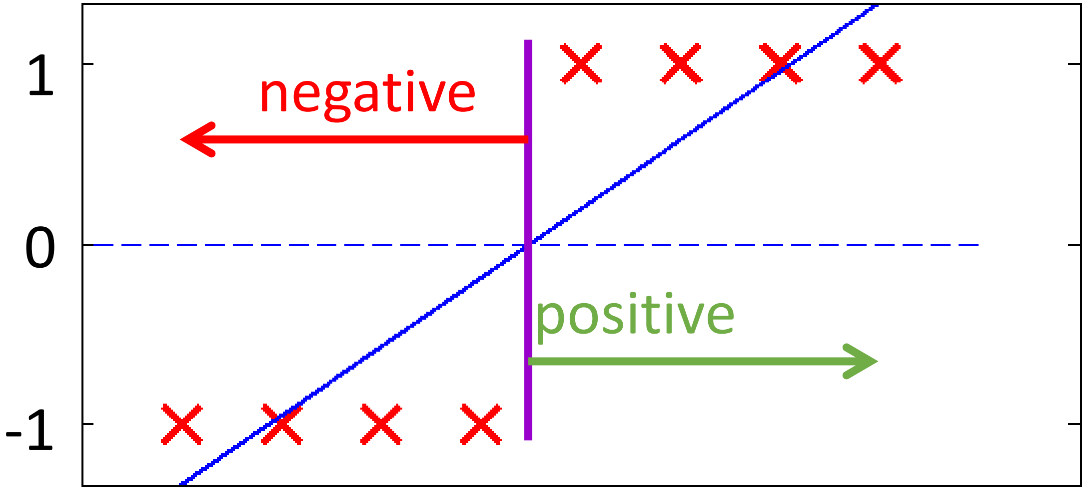
- U slučaju binarnih varijabli klase možemo obeležiti sa $y \in \{-1, +1\}$, a $\{-1, +1\} \in \mathbb{R}$

- Dakle, možemo primeniti linearnu regresiju da odredimo θ tako da je $\theta^T x_n \approx y_n$, gde $y_n \in \{-1, +1\}$

Možemo li primeniti linearnu regresiju?

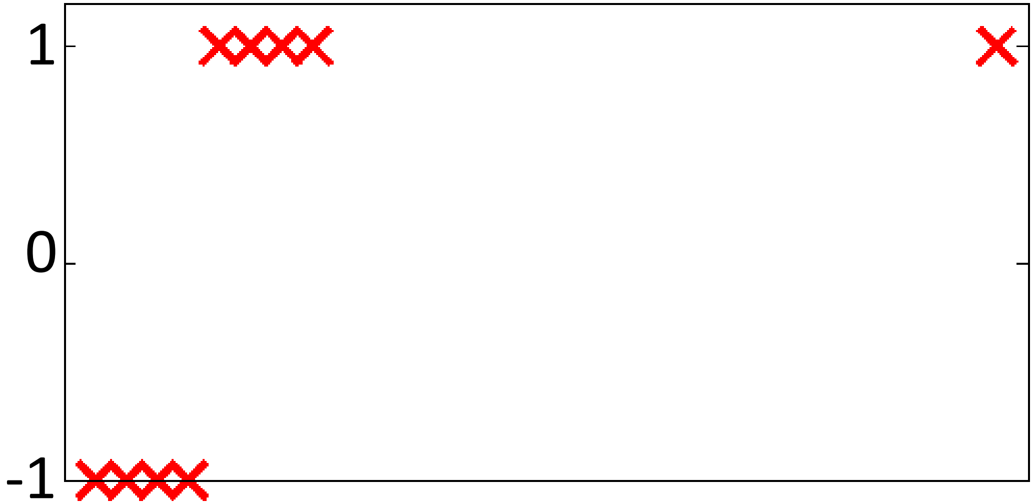


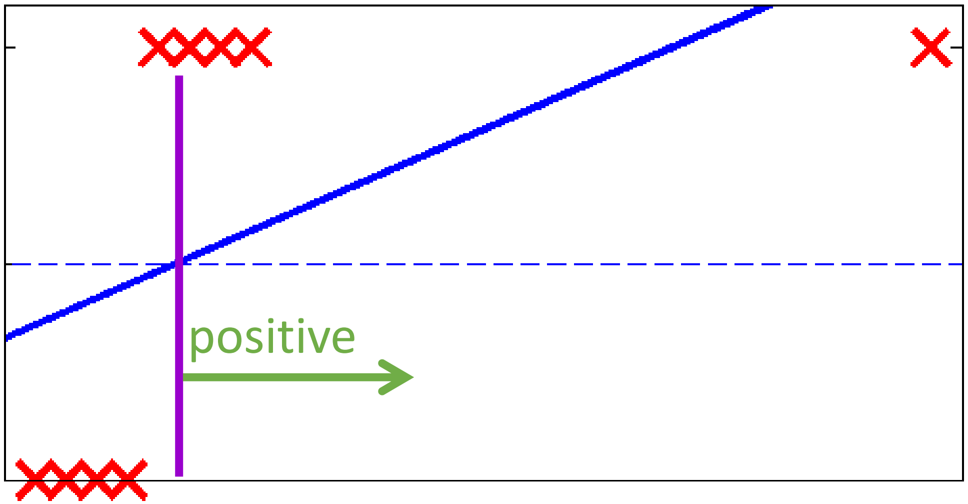
Perfektno balasnirani podaci



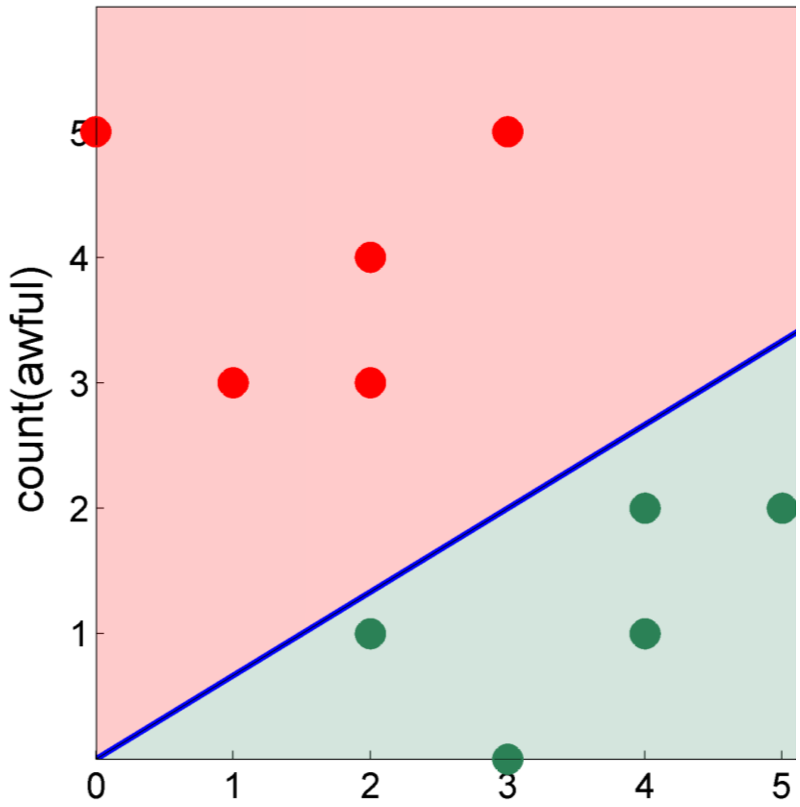
$$y = sign(\theta^T x)$$

Možemo li primeniti linearnu regresiju?





$$y = sign(\theta^T x)$$

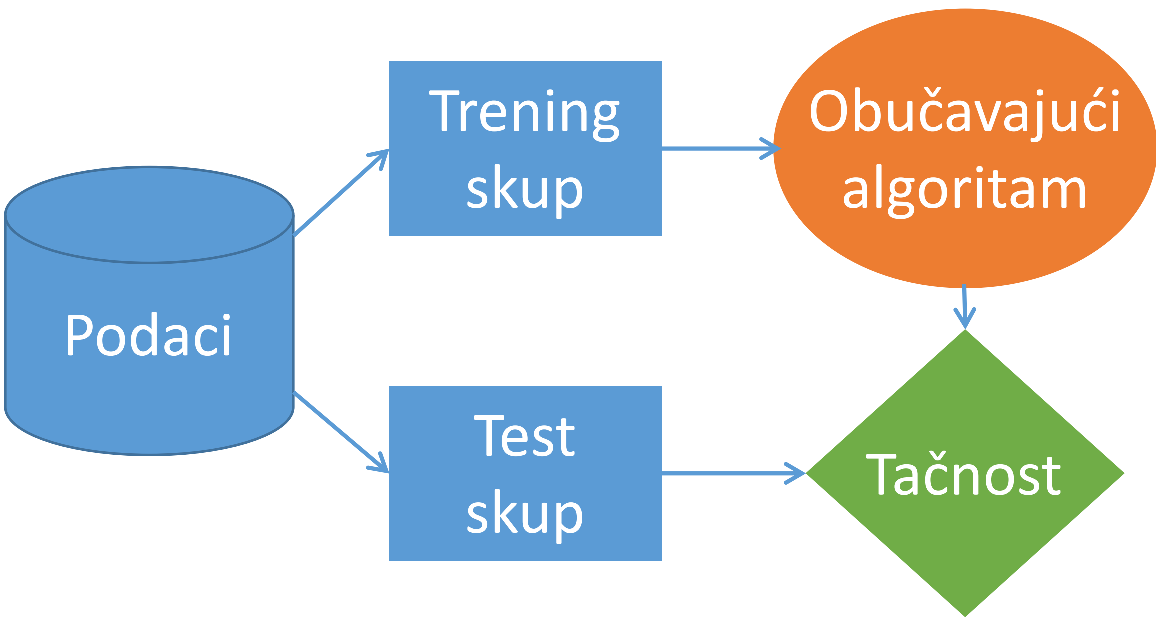


Linearni klasifikator

Primer: Klasifikacija recenzija restorana na pozitivne i negativne

Linearan klasifikator

- Svakoj reči iz recenzija dodeljijemo težinski koeficijent
- Koeficijent izražava koliko pozitivnog/negativnog uticaja ima data reč
- Treniranje modela: određivanje koeficijenata θ na osnovu trening skupa



Reč	θ
good	1.0
great	1.2
...	...

Linearni klasifikator

- Izlaz je težinska suma ulaza

Reč	θ
…	…

Recenzija

Jednostavan linearan klasifikator

$Score(x)$ = težinska suma broja pojave reči u rečenici

ako je $Score(x) > 0$:

$$\hat{y} = 1$$

u suprotnom:

$$\hat{y} = 0$$

$$x_1 = 1$$
$$x_2 = count(\text{"}awesome\text{"})$$
$$x_3 = count(\text{awful})$$
$$x_4 = count(\text{"}good\text{"})$$
…

$$\hat{y}^{(i)} = sign\left(score\left(x^{(i)}\right)\right)$$

$$score\left(x^{(i)}\right) = \sum_{d=1}^{D} \theta_d x_d^{(i)} = \theta x^{(i)}$$

Linearan klasifikator

Reč	$\boldsymbol{\theta}$
good	1.0
great	1.2
awsome	1.7
bad	-1.0
terrible	-2.1
awful	-3.3
restaurant, the, we, where,…	0.0
…	…

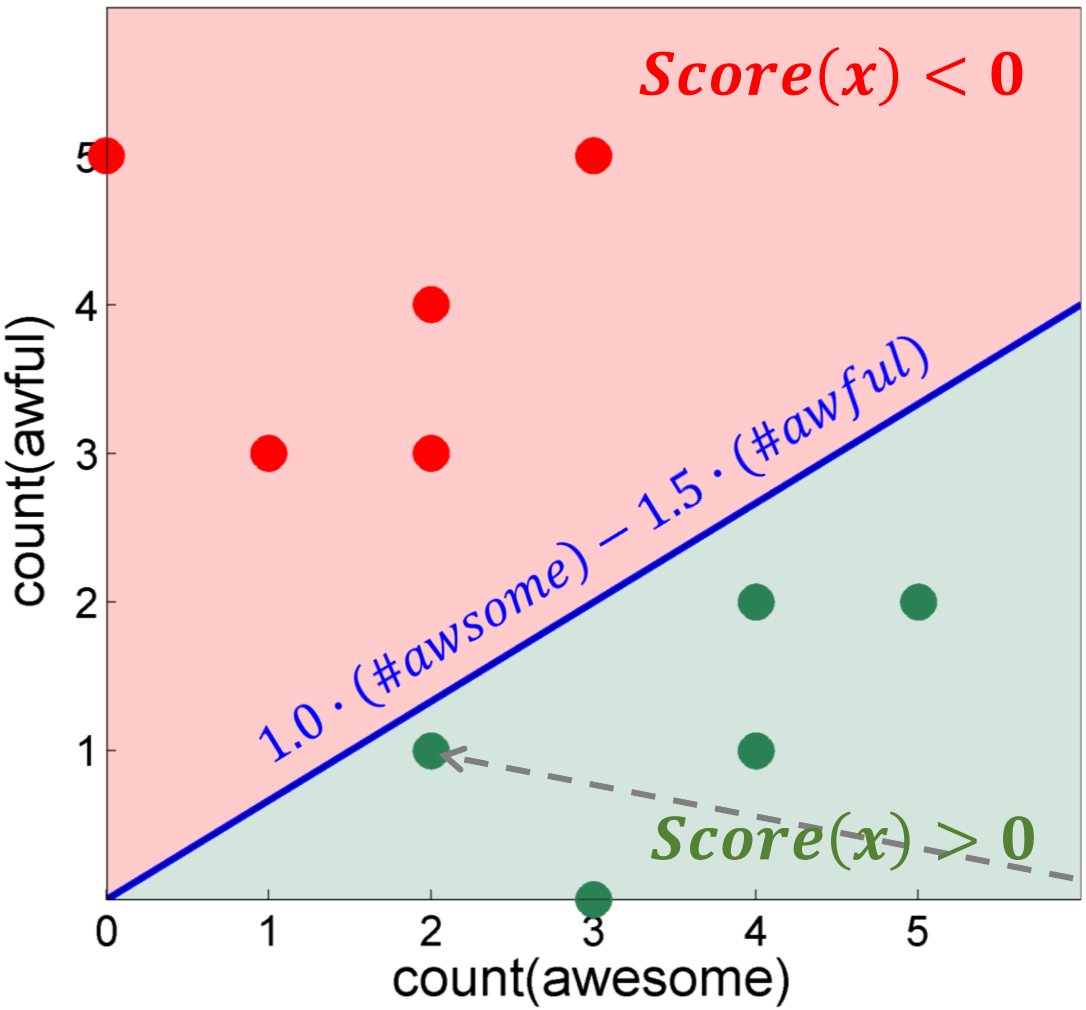
Ulaz $x^{(i)}$:

Sushi was great, the food was awsome,
But the service was terrible.

$$\theta^T x^{(i)} = 1.2 + 1.7 - 2.1$$
$$= 0.8 > 0$$

$$\Longrightarrow \hat{y}^{(i)} = 1$$

Granica odluke



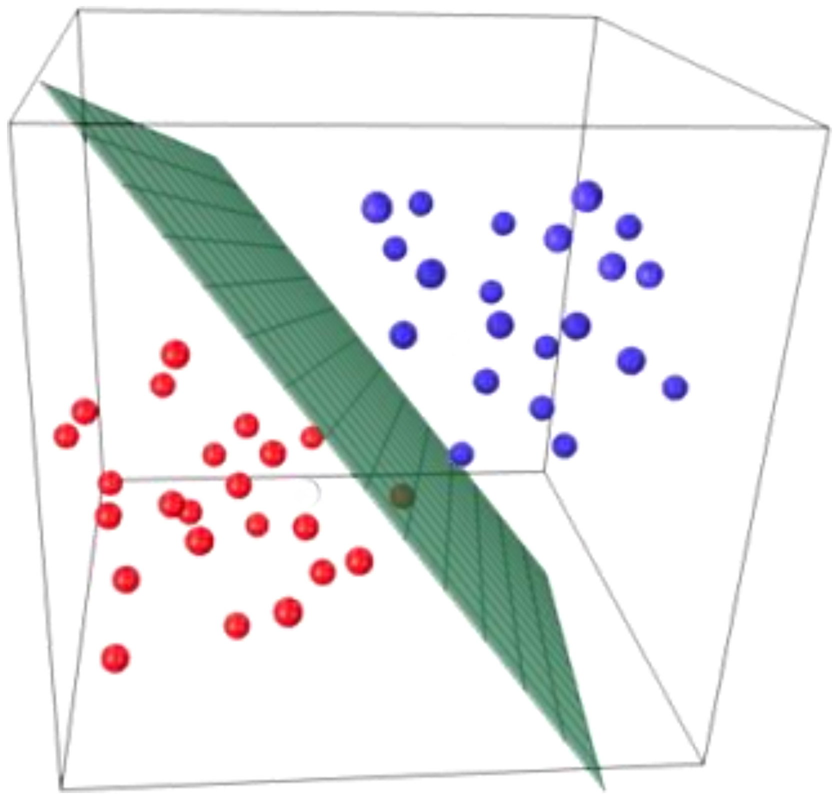
Reč	θ
awesome	1.0
awful	-1.5

$$Score(x) = 1.0 \cdot count(awsome) - 1.5 \cdot count(awful)$$

Sushi was awesome, the food was awsome, but the service was awful.

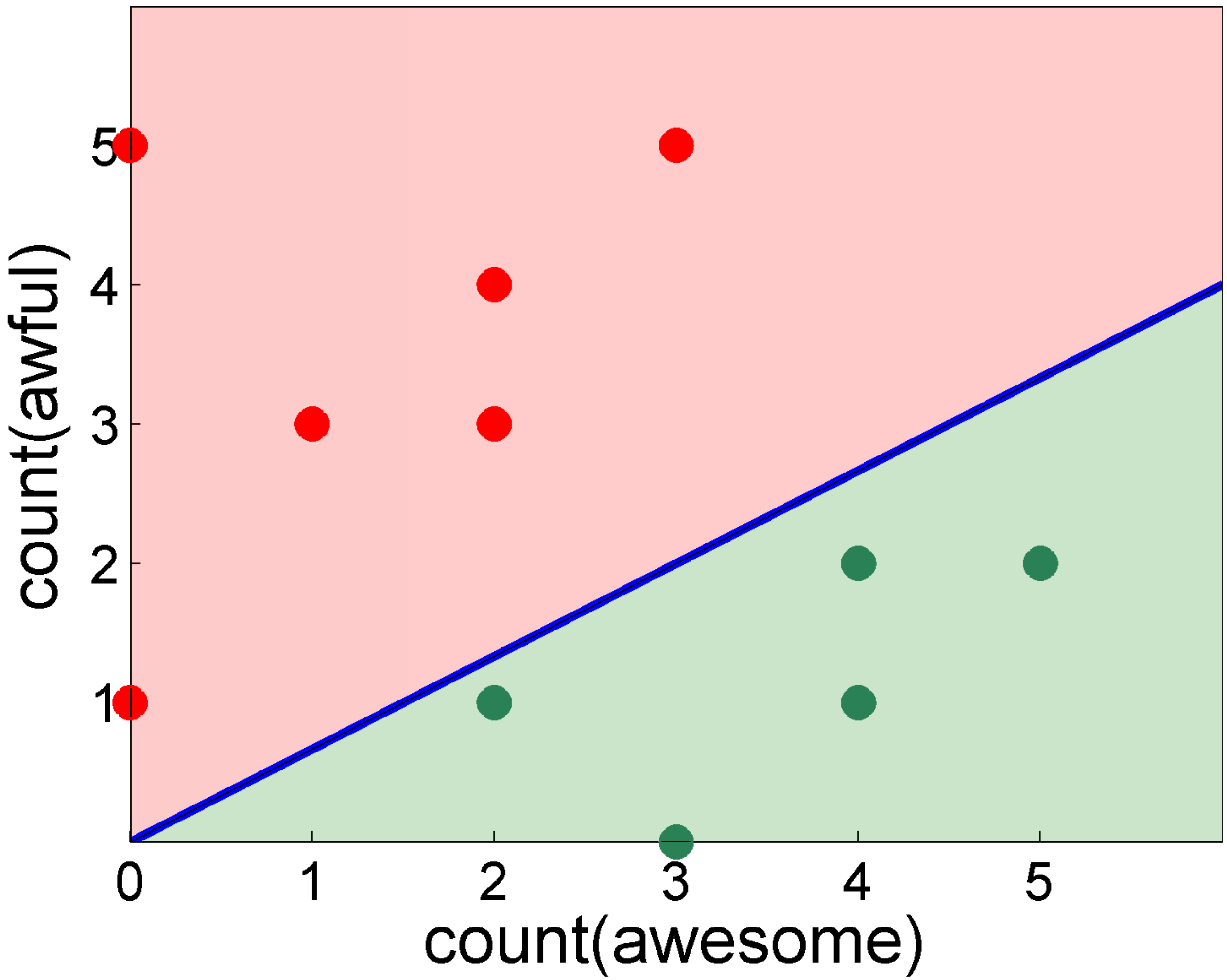
Granica odluke

- U slučaju da imamo *D* obeležja sa ne-nula koeficijentima, granica odluke bi bila hiperravan
- Npr. za D=3:



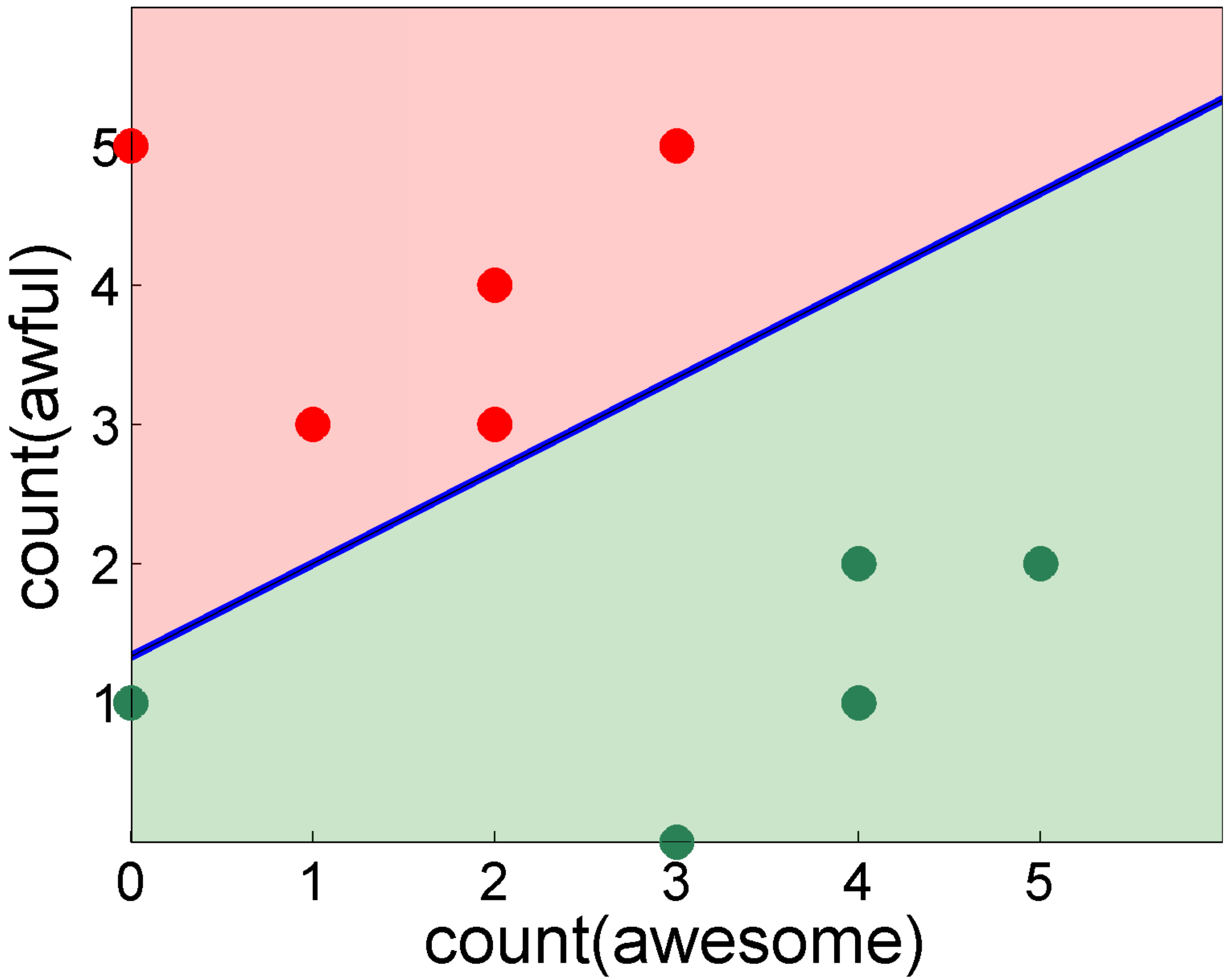
Efekat koeficijenata na granicu odluke

$$score(x) = 1.0 \cdot count(awsome) - 1.5 \cdot count(awful)$$



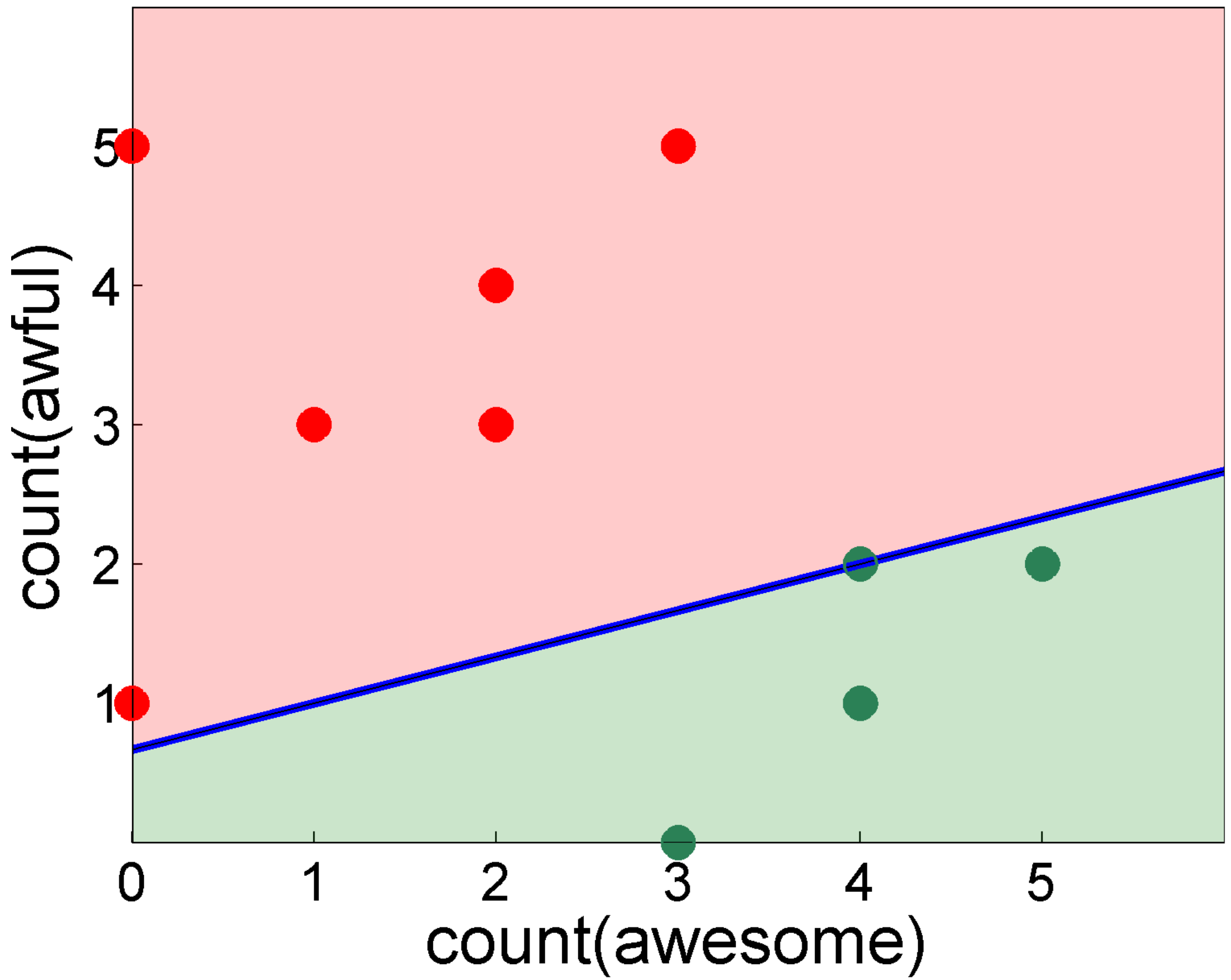
Efekat koeficijenata na granicu odluke

$$score(x) = \color{red}{2.0} + 1.0 \cdot count(awsome) - 1.5 \cdot count(awful)$$



Efekat koeficijenata na granicu odluke

$$score(x) = 2.0 + 1.0 \cdot count(awsome) - 3.0 \cdot count(awful)$$



Verovatnoće klasa

- Prilikom klasifikacije nećemo predviđati samo $\hat{y} = 1$ ili $\hat{y} = 0$
- Klasifikator će nam kao izlaz dati procenu verovatnoće $P(y^{(i)} = 1|x^{(i)})$
- Estimacija $P(y^{(i)} = 1|x^{(i)})$ pobojšava interpretabilnost modela – govori koliko smo sigurni u datu predikciju

The sushi & everything else was awesome

$\hat{y} = 1$ sa verovatnoćom 1

The sushi was good, the service was OK

$\hat{y} = 1$ sa verovatnoćom 0.5

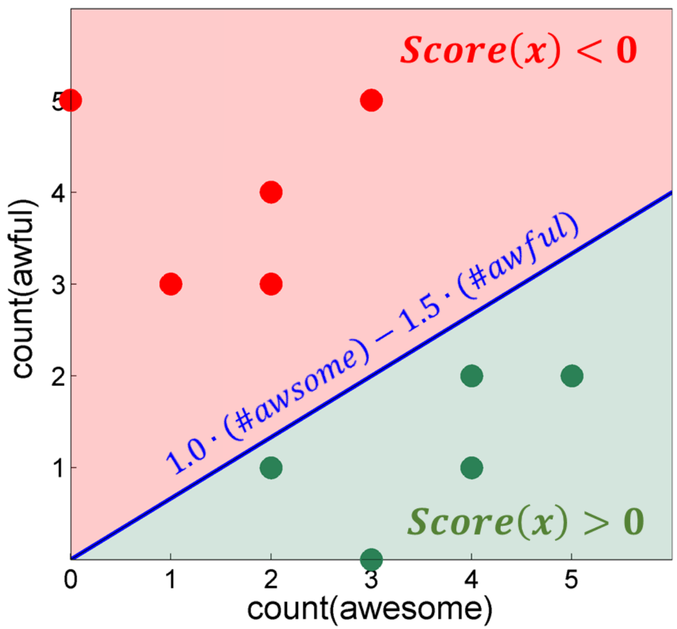
ako je $P(y^{(i)} = 1|x^{(i)}) > 0.5$:

$\hat{y} = 1$

u suprotnom:

$\hat{y} = 0$

Generalized linear models (GLM)



$-\infty$ $\quad score\left(x^{(i)}\right) = \theta x^{(i)} \quad$ ∞

$\hat{y}^{(i)} = 0$

$P\left(y^{(i)} = 1 | x^{(i)}\right) = 0$

$\hat{y}^{(i)} = 1$

$P\left(y^{(i)} = 1 | x^{(i)}\right) = 1$

$score\left(x^{(i)}\right) = 0$

$P\left(y^{(i)} = 1 | x^{(i)}\right) = 0.5$

- Definisaćemo *link function g:*

$$h_\theta(x) = P\left(y = 1 | x^{(i)}, \theta\right) = g\left(\theta x^{(i)}\right)$$

- Ovakav model se zove *Generalized Linear Model* (GLM)

Da li je $h_\theta(x)$ *stvarna* verovatnoća?

- Da li $h_\theta(x)$ možemo tretirati kao *stvarnu* verovatnoću $P\left(y^{(i)} = 1|x^{(i)}\right)$?
- Podaci nam ne govore ništa o verovatnoći – dobijamo (x, y) sa *binarnim* y koje može da sadrži šum:

$$P(y|x) = \begin{cases} f(x) & \text{za } y = +1 \\ 1 - f(x) & \text{za } y = -1 \end{cases}$$

- Ono što mi pokušavamo da naučimo je $f(x)$, bez obzira na to što nam je poznato samo y (generisano od strane $f(x)$)
- Zato o $h(x)$ možemo zaista razmišljati kao o stvarnoj verovatnoći – ne samo da je u opsegu $[0,1]$ već primeri koji su nam dati imaju suštinsku probabilističku interpretaciju