Announcements

Quiz: next week after the lecture, will be live until the end of the day.

4 questions, multiple choice, 20 mins

2 marks (0.5/question)

Will be on moodle (you can take it essentially at home)

Coursework: MATLAB

Final project

On cross entropy loss **fix moodle issues scheduling issues.**

INM431: Machine Learning

Naïve Bayes

Pranava Madhyastha (pranava.madhyastha@city.ac.uk)

Based on materials by Artur S. d'Avila Garcez + Bishop's book + Murphy's book

Probability estimates

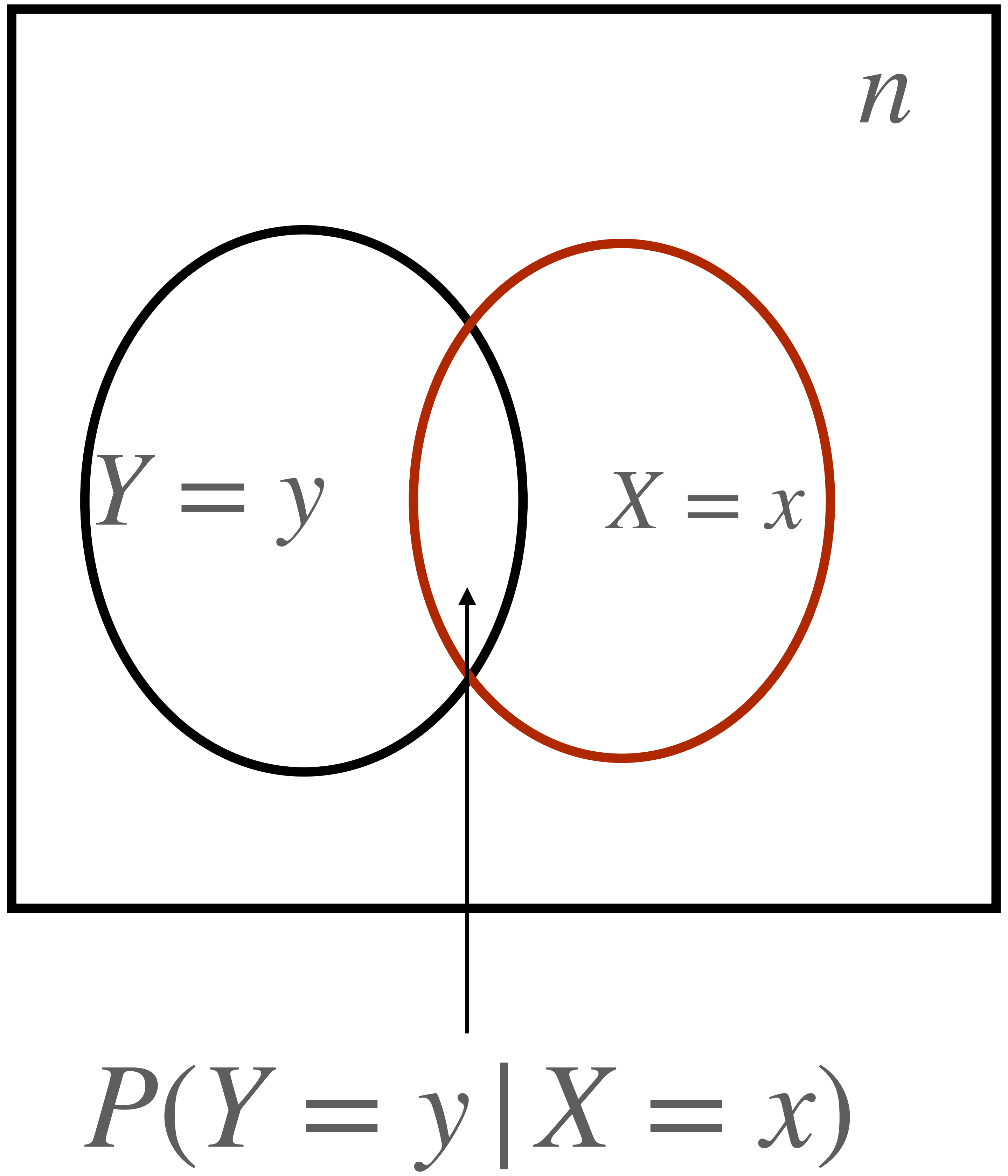
Let's look at this: $P(X = x) = \dfrac{\sum_{i=1}^{n} C(X_i = x)}{n}$

What is the value of $P(Y = y \,|\, X = x) = ?$

Probability estimates

Let's look at this: $P(X = x) = \dfrac{\sum_{i=1}^{n} C(X_i = x)}{n}$

What is the value of $P(Y = y \mid X = x) = ?$

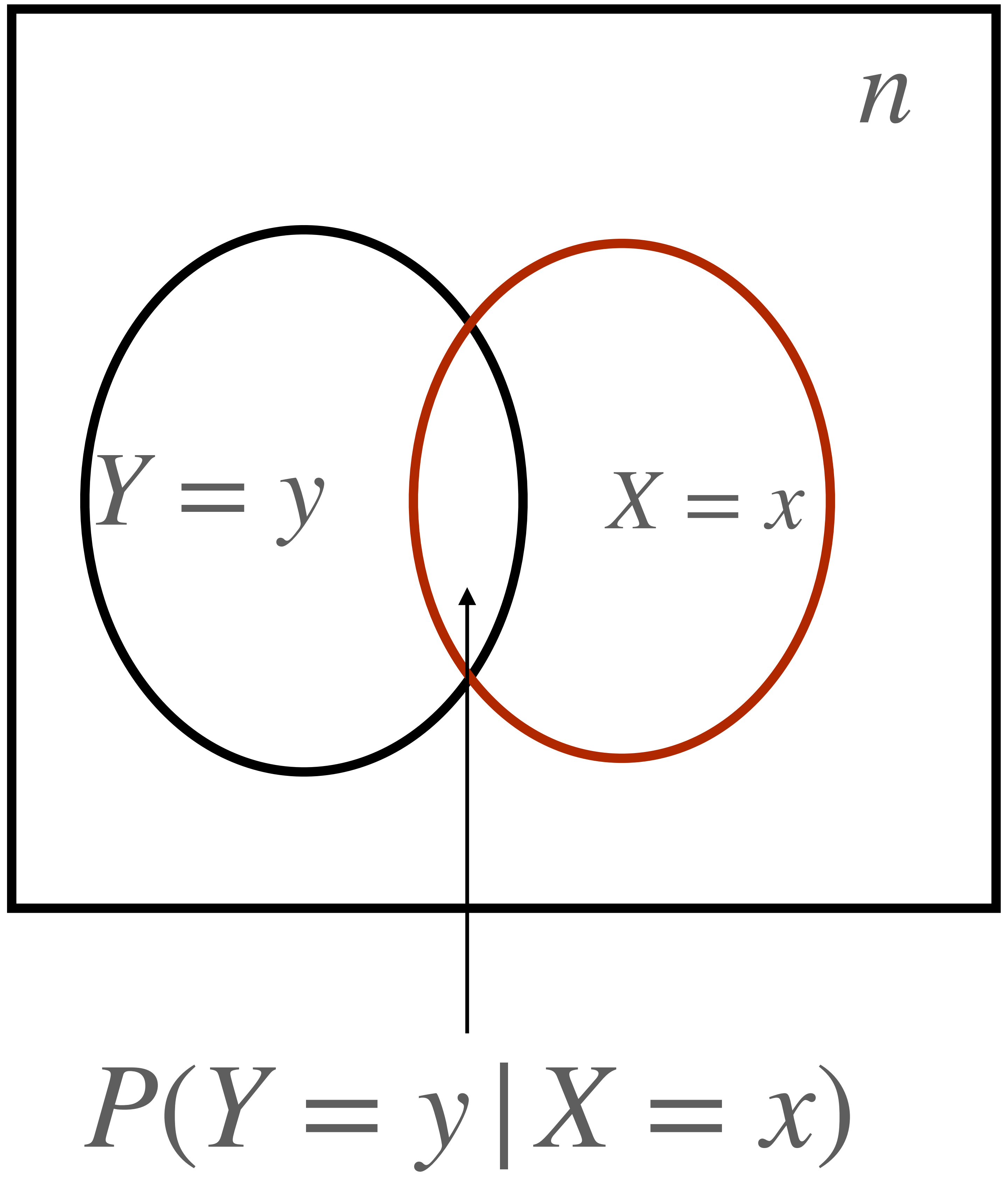


Probability estimates

$$P(Y=y\,|\,X=x) = \frac{P(Y=y, X=x)}{P(X=x)}$$

$$= \frac{\frac{\sum_{i=1}^{n} C(X_i=x, Y_i=y)}{n}}{\frac{\sum_{n=1}^{n} C(X_i=x)}{n}}$$

$$= \frac{\sum_{i=1}^{n} C(X_i=x, Y_i=y)}{\sum_{n=1}^{n} C(X_i=x)}$$



d-dimensional case

$P(Y = y \mid X_1 = x_1, \cdots, X_d = x_d)$ = ?

Every single feature has to be exactly same!

- Deterministic

- Not a great algorithm, most of the times it would be useless.

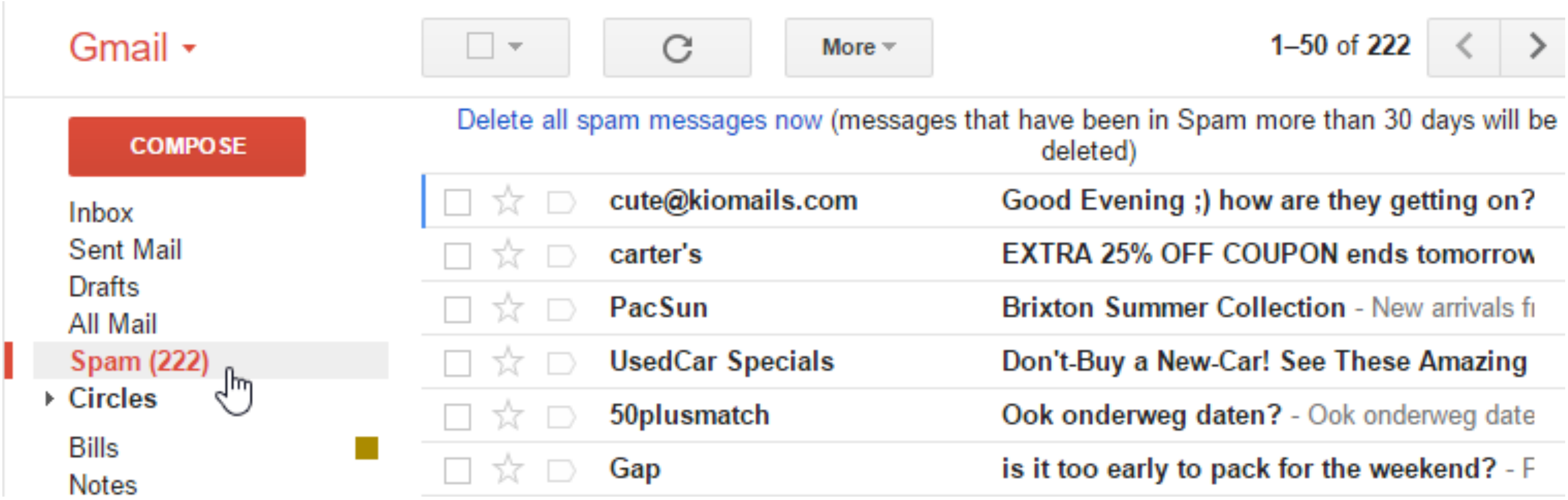
Naïve Bayes

Let's use Bayes' theorem:

$$P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)}$$

The assumption: Each feature is independent from each other given the label!

Naïve Bayes


Gmail
More
1–50 of 222
COMPOSE
Inbox
Sent Mail
Drafts
All Mail
Spam (222)
Circles
Bills
Notes
Delete all spam messages now (messages that have been in Spam more than 30 days will be deleted)
cute@kiomails.com
Good Evening ;) how are they getting on?
carter's
EXTRA 25% OFF COUPON ends tomorrow
PacSun
Brixton Summer Collection - New arrivals fi
UsedCar Specials
Don't-Buy a New-Car! See These Amazing
50plusmatch
Ook onderweg daten? - Ook onderweg date
Gap
is it too early to pack for the weekend? - F

Naïve Bayes

The assumption:

$$P(X = x \mid Y = y) = \prod_{j=1}^{d} P(|X|_j = x_j \mid Y = y)$$

Each feature is independent from each other given the label!

Big trade off, but it is now computational tractable!

Naïve Bayes

The assumption:

$$P(X = x \,|\, Y = y) = \prod_{j=1}^{d} P(\,|X|_j = x_j \,|\, Y = y)$$

A good place where it fits: when there is a good causal relationship.

Typically, in a clinical setting.

Naïve Bayes

Optimal classifier:

$$\text{argmax}_y P(y \mid \overrightarrow{X}) = \text{argmax}_y \frac{P(\overrightarrow{X} \mid y)P(y)}{Z}$$

$$= \text{argmax}_y \ P(y)\prod_{j=1}^{d} P(\,|X|_j \mid y)$$

$$= \text{argmax}_y \ \log P(y) + \sum_{j=1}^{d} \log P(\,|X|_j \mid y)$$

Naïve Bayes family of classifiers

A class prior may be calculated by assuming equiprobable classes:

$$\text{prior} = \frac{1}{|\text{classes}|}$$

or by calculating an estimate for the class probability from the training set:

$$\text{class prior} = \frac{(\text{number of samples in the class})}{(\text{total number of samples})}$$

Naïve Bayes is a family of classifiers because it applies to any distribution. One can have a Gaussian Naïve Bayes, a Bernoulli naïve Bayes, a multinomial Naïve Bayes, etc,

Despite the naïve conditional independence assumption, naïve Bayes classifiers can be surprisingly efficient on various datasets

Example: Flu=Yes/No?

chills	runny nose	headache	fever	Flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

chills	runny nose	headache	fever	Flu?
Y	N	Mild	N	?

Example: Flu=Yes/No?

Prior: P(flu=Y) = 5/8, P(flu=N) = 3/8

Likelihoods:

P(chills=Y|flu=Y) = 3/5 P(chills=N|flu=Y) = 2/5

P(runny nose=Y|flu=Y) = 4/5 P(runny nose=N|flu=Y) = 1/5

P(headache=mild|flu=Y) = 2/5 P(headache=no|flu=Y) = 1/5 P(headache=strong|flu=Y) = 2/5

P(fever=Y|flu=Y) = 4/5 P(fever=N|flu=Y) = 1/5

chills	runny nose	headache	fever	Flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

chills	runny nose	headache	fever	Flu?
Y	N	Mild	N	?

Example: Flu=Yes/No?

chills	runny nose	headache	fever	Flu?
Y	N	Mild	N	?

Now, we have to compute:

Posterior 1: $P(flu = Y \mid chills = Y, runny\ nose = N, headache = mild, fever = N)$

P(flu=Y))P(chills=Y | flu=Y)P(runny=Y | flu=Y)P(headache=mild | flu=Y)P(fever=N | $flu = Y$)

= 0.006

v/s

Posterior 2: $P(flu = N \mid chills = Y, runny\ nose = N, headache = mild, fever = N)$

P(flu=N))P(chills=Y | flu=N)P(runny=Y | flu=N)P(headache=mild | flu=N)P(fever=N | $flu = N$)

= 0.0185

$$\arg\max\{Y = 0.006, N = 0.0185\}$$

= No FLU!

Naïve Bayes

The assumption:

$$P(X = x \mid Y = y) = \prod_{j=1}^{d} P(|X|_j = x_j \mid Y = y)$$

When is it good/when is it bad?

It’s a non-parametric model!

Multinomial distribution: case of email

$x_j \in \{0, \cdots, m\}$

The number of times a word appears in the email.

$$m = \sum_{j=1}^{d} x_j$$

$$\begin{array}{r} \text{a} \\ \vdots \\ \text{Sincerely} \\ \vdots \end{array} \begin{pmatrix} 25 \\ \vdots \\ 1 \\ \vdots \end{pmatrix}$$

Multinomial distribution: case of email

$$P(\overrightarrow{x} \mid m, y = \text{spam}) = ?$$

$$= \frac{m!}{x_1!, \cdots, x_d!} \prod_{j=1}^{d} (\theta_{j,\text{spam}})^{x_j}$$

$$\begin{array}{r} \text{a} \\ \vdots \\ \text{Sincerely} \\ \vdots \end{array} \begin{pmatrix} 25 \\ \vdots \\ 1 \end{pmatrix} \\ \quad\quad\quad\quad\quad\quad \vdots$$

Multinomial distribution: case of email

$$P(\overrightarrow{x} \,|\, m, y = \text{spam}) = \frac{m!}{x_1!, \cdots, x_d!} \prod_{j=1}^{d} \theta_{j,\text{spam}}^{x_j}$$

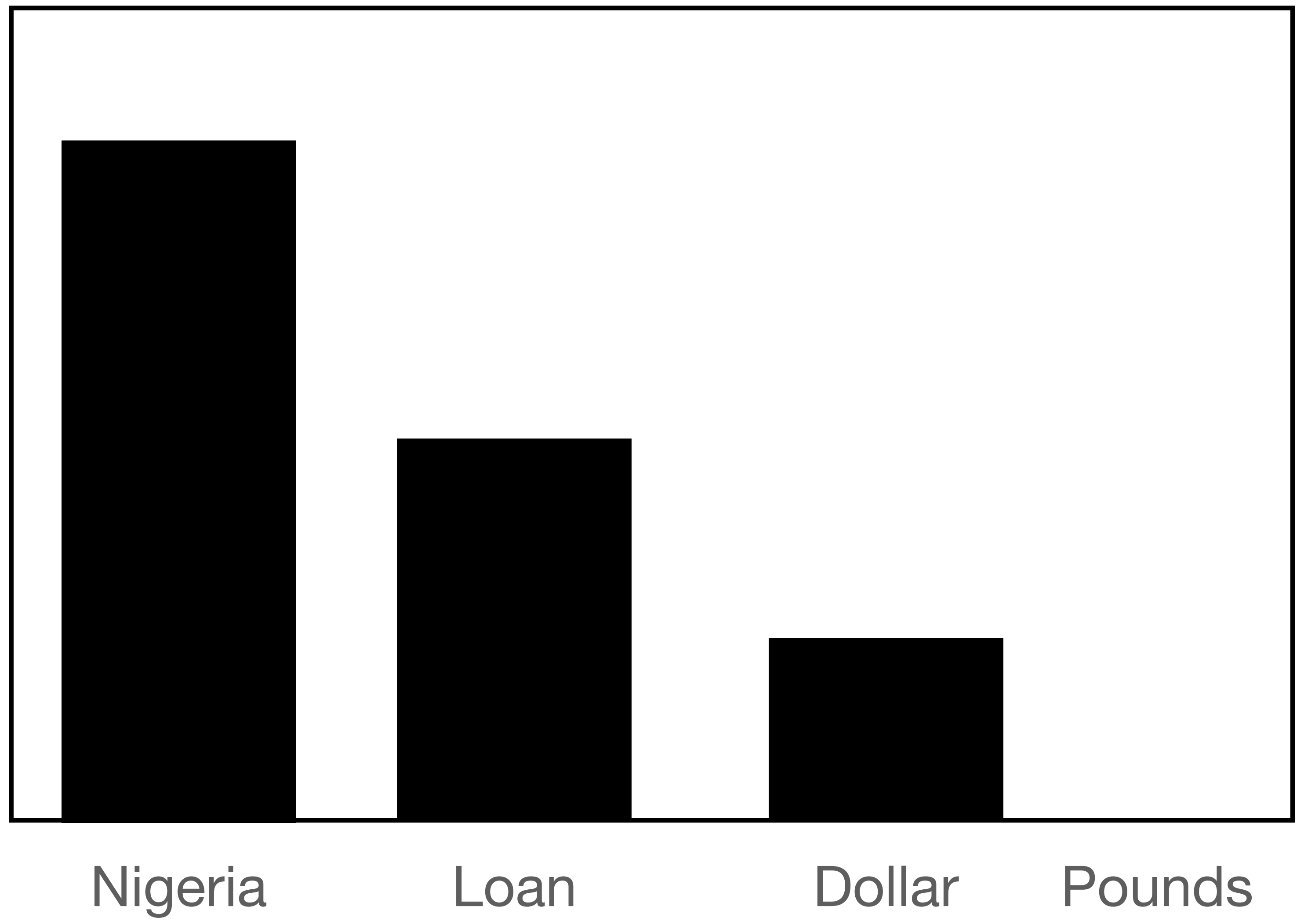
$$\theta_{j,\text{spam}} = \frac{\sum_{i=1}^{n} C(y_i = \text{spam}) x_{ij}}{\sum_{i=1}^{n} C(y_i = \text{spam}) (\sum_{j=1}^{d} x_{ij})}$$

$$\begin{array}{r c} \begin{array}{r} \text{a} \\ \\ \vdots \\ \\ \text{Sincerely} \end{array} & \begin{pmatrix} 25 \\ \\ \vdots \\ \\ 1 \end{pmatrix} \\ \vdots & \vdots \end{array}$$

The concept of smoothing

$$P(X = x \mid Y = y) = \prod_{j=1}^{d} P(|X|_j = x_j \mid Y = y)$$

What if some of the $|X|_j = 0$?



The concept of smoothing

Let $x = (x_1, \cdots, x_d)$ be observation from a multinomial distribution with N trials (x_i is the number of times outcome i is observed)

A smoothed version of each x_i is given by $\dfrac{(x_i + 1)}{(N + d)}$

The Bayesian analogue: The resulting estimate will be between the empirical probability (or relative frequency) $\dfrac{x_i}{N}$ and the uniform probability $\dfrac{1}{d}$

Add one smoothing/Laplace smoothing

Let $x = (x_1, \cdots, x_d)$ be observation from a multinomial distribution with N trials (x_i is the number of times outcome i is observed)

A smoothed version of each x_i is given by $\dfrac{(x_i + 1)}{(N + d)}$

The Bayesian analogue: The resulting estimate will be between the empirical probability (or relative frequency) $\dfrac{x_i}{N}$ and the uniform probability $\dfrac{1}{d}$

Our parameter with smoothing

$$P(\overrightarrow{x} \mid m, y = \text{spam}) = \frac{m!}{x_1!, \cdots, x_d!} \prod_{j=1}^{d} \theta_{j,\text{spam}}^{x_j}$$

$$\theta_{j,\text{spam}} = \frac{\sum_{i=1}^{n} C(y_i = \text{spam}) x_{ij} + 1}{\sum_{i=1}^{n} C(y_i = \text{spam}) (\sum_{j=1}^{d} x_j) + d}$$

$$\begin{array}{c} \text{a} \\ \vdots \\ \text{Sincerely} \\ \vdots \end{array} \begin{pmatrix} 25 \\ \vdots \\ 1 \end{pmatrix} \begin{array}{c} \\ \\ \\ \vdots \end{array}$$

Our parameter with smoothing

$$P(\overrightarrow{x} \mid m, y = \text{spam}) = \frac{m!}{x_1!, \cdots, x_d!} \prod_{j=1}^{d} \theta_{j,\text{spam}}^{x_j}$$

$$\theta_{j,\text{spam}} = \frac{\sum_{i=1}^{n} C(y_i = \text{spam}) x_{ij} + 1}{\sum_{i=1}^{n} C(y_i = \text{spam}) (\sum_{j=1}^{d} x_j) + d}$$

$$\begin{array}{r} \text{a} \\ \vdots \\ \text{Sincerely} \\ \vdots \end{array} \begin{array}{c} \left(\begin{array}{c} 25 \\ \vdots \\ 1 \end{array} \right) \\ \vdots \end{array}$$

Continuous and discrete data

Since we have the conditional independence assumption in Naive Bayes, we can in fact mix variables.

We can compute the likelihoods of binary variables using a Bernoulli distribution, and compute the likelihoods of the continuous variables with a Gaussian.

Gaussian Naïve Bayes

When the data is continuous: $x_j \in \mathscr{R}$

Priors are calculated as before

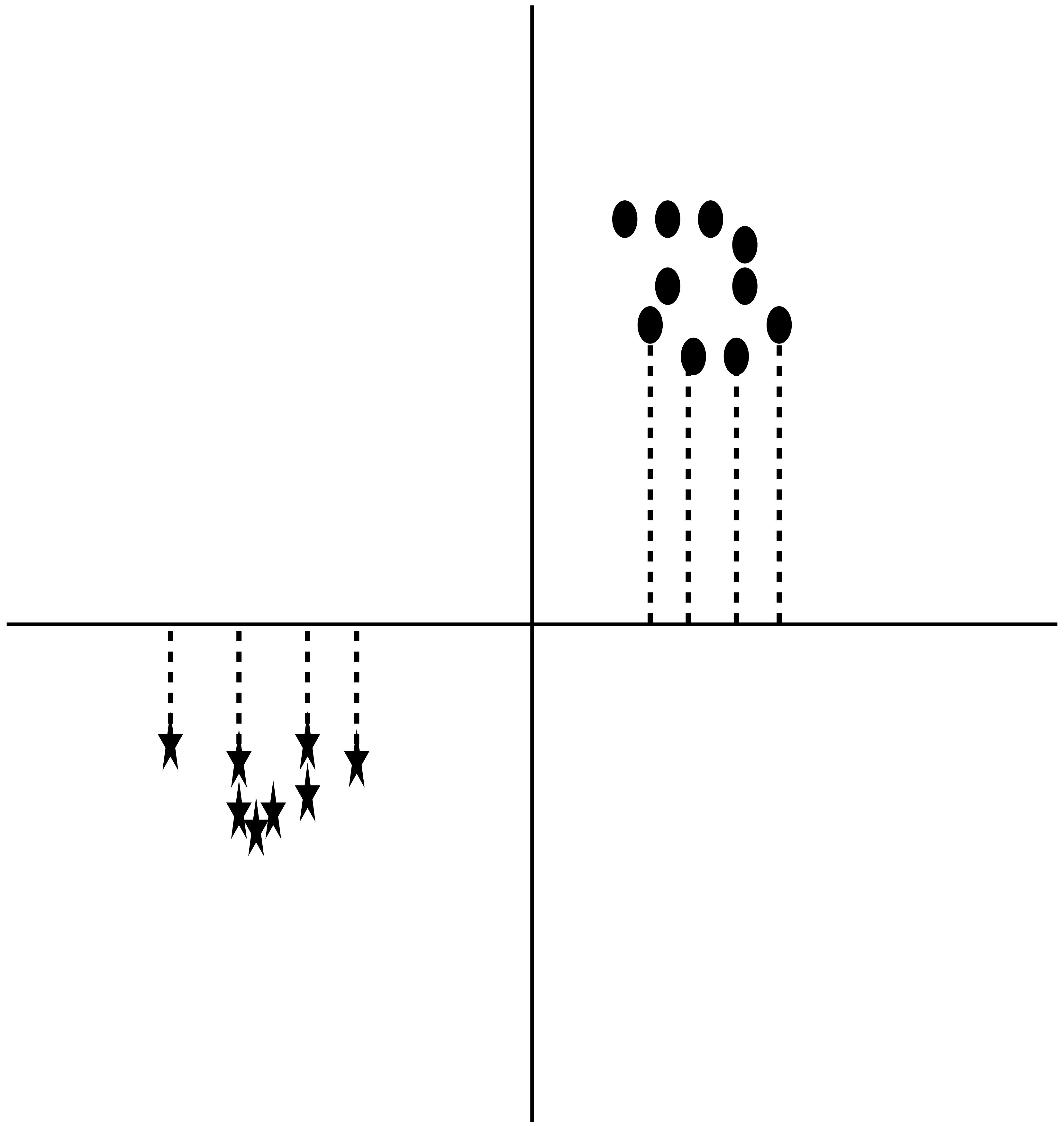
Likelihoods are calculated from the training set by finding mean and variance for each attribute given a class:

$$P(x_j \,|\, y = C_k) = \mathscr{N}(\mu_{jk}, \sigma_{jk}^2)$$

Posteriors are calculated by multiplying priors and likelihoods, producing a Gaussian for each class:

Given test data, $P(y = C_k \,|\, x_j)$ can now be calculated

in the case of continuous variable x taking value x_j



Back to the multinomial

Assume that we have two classes: {-1,+1}

$$P(Y=+1\,|\,x) > P(Y=-1\,|\,x)$$

$$P(+1)P(\overrightarrow{x}\,|+1) > P(-1)P(\overrightarrow{x}\,|-1)$$

$$\frac{m!}{x_1!,\ldots,x_d!}\prod_{j=1}^{d}\theta_{j,+1}^{x_j} > \frac{m!}{x_1!,\ldots,x_d!}\prod_{j=1}^{d}\theta_{j,-1}^{x_j}$$

Back to the multinomial

Assume that we have two classes: {-1,+1}

$$P(Y = +1 \,|\, x) > P(Y = -1 \,|\, x)$$

$$P(+1)P(\overrightarrow{x} \,|\, +1) > P(-1)P(\overrightarrow{x} \,|\, -1)$$

$$P(+1)\frac{m!}{x_1!, \cdots, x_d!}\prod_{j=1}^{d}\theta_{j,+1}^{x_j} > P(-1)\frac{m!}{x_1!, \cdots, x_d!}\prod_{j=1}^{d}\theta_{j,-1}^{x_j}$$

Lets take the logs

$$\log P(+1) \sum_{j=1}^{d} x_j \log \theta_{j,+1} > P(-1) \sum_{j=1}^{d} x_j \log \theta_{j,-1}$$

Rearranging

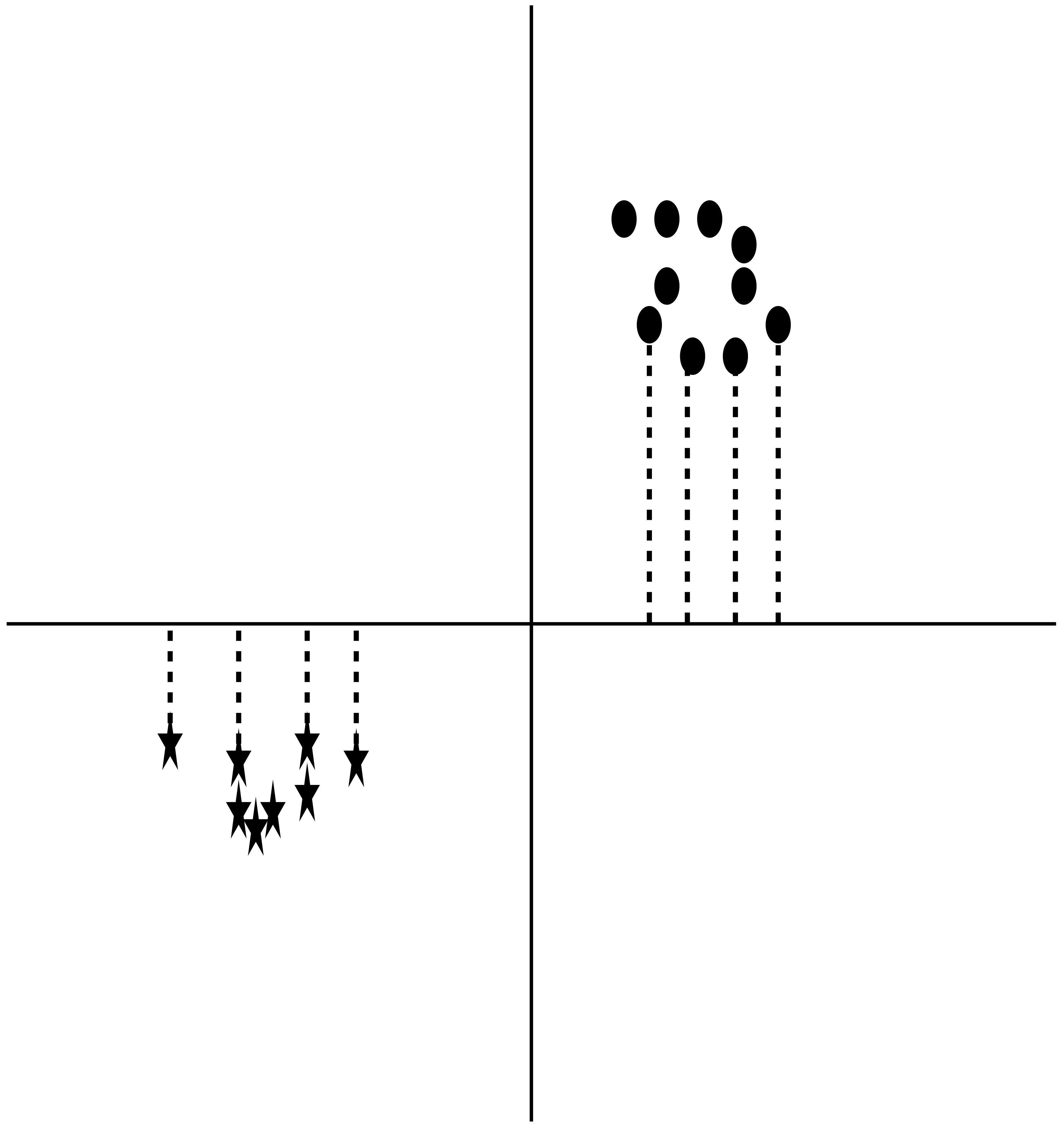
$$\boxed{\log P(+1) - \log P(-1)} + \sum_{j=1}^{d} x_j \boxed{(\log \theta_{j,+1} - \theta_{j,-1})} > 0$$

$b + w^\top x > 0 \leftarrow$ linear classifier!

Rearranging

$$\log P(+1) - \log P(-1) + \sum_{j=1}^{d} x_j(\log \theta_{j,+1} - \theta_{j,-1}) > 0$$

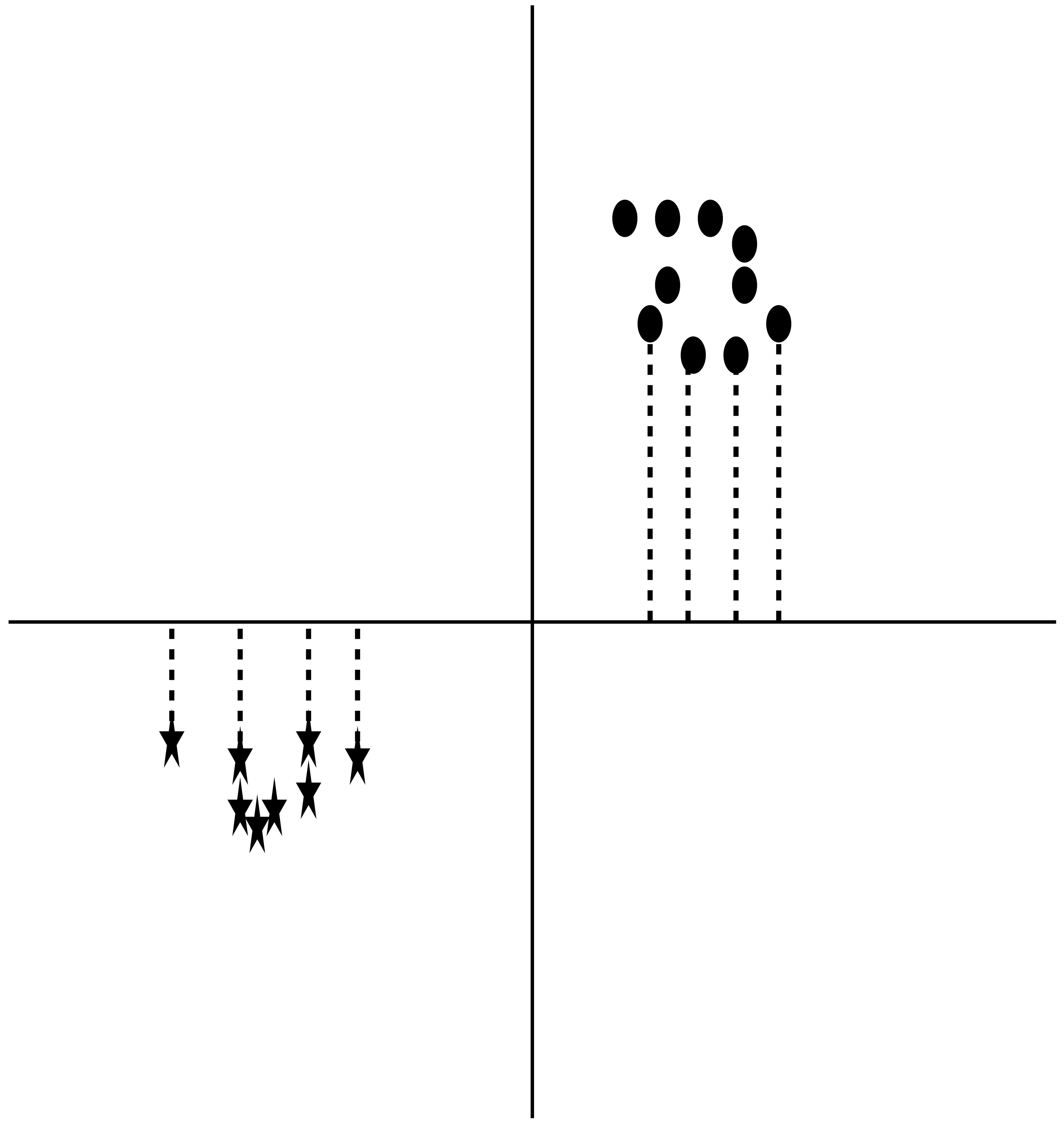
$b + w^\top x > 0 \leftarrow$ linear classifier!



In the case of Gaussian

$$P(y \mid \overrightarrow{x}) = \frac{1}{1 + exp(-w^{\top}\frac{\overrightarrow{x}y}{Z})}$$

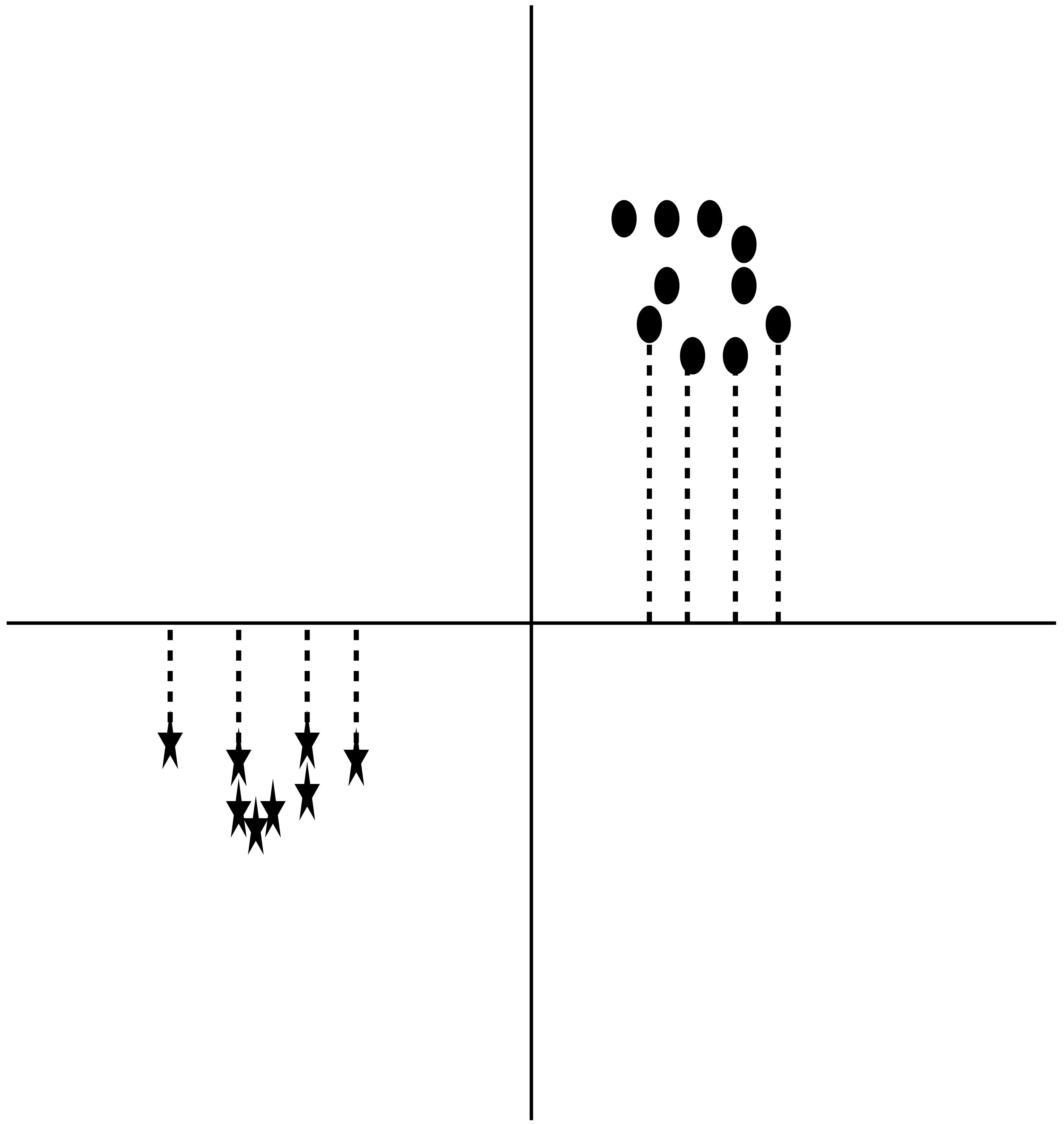
where, $y = \pm 1$



Generative v/s discriminative approach

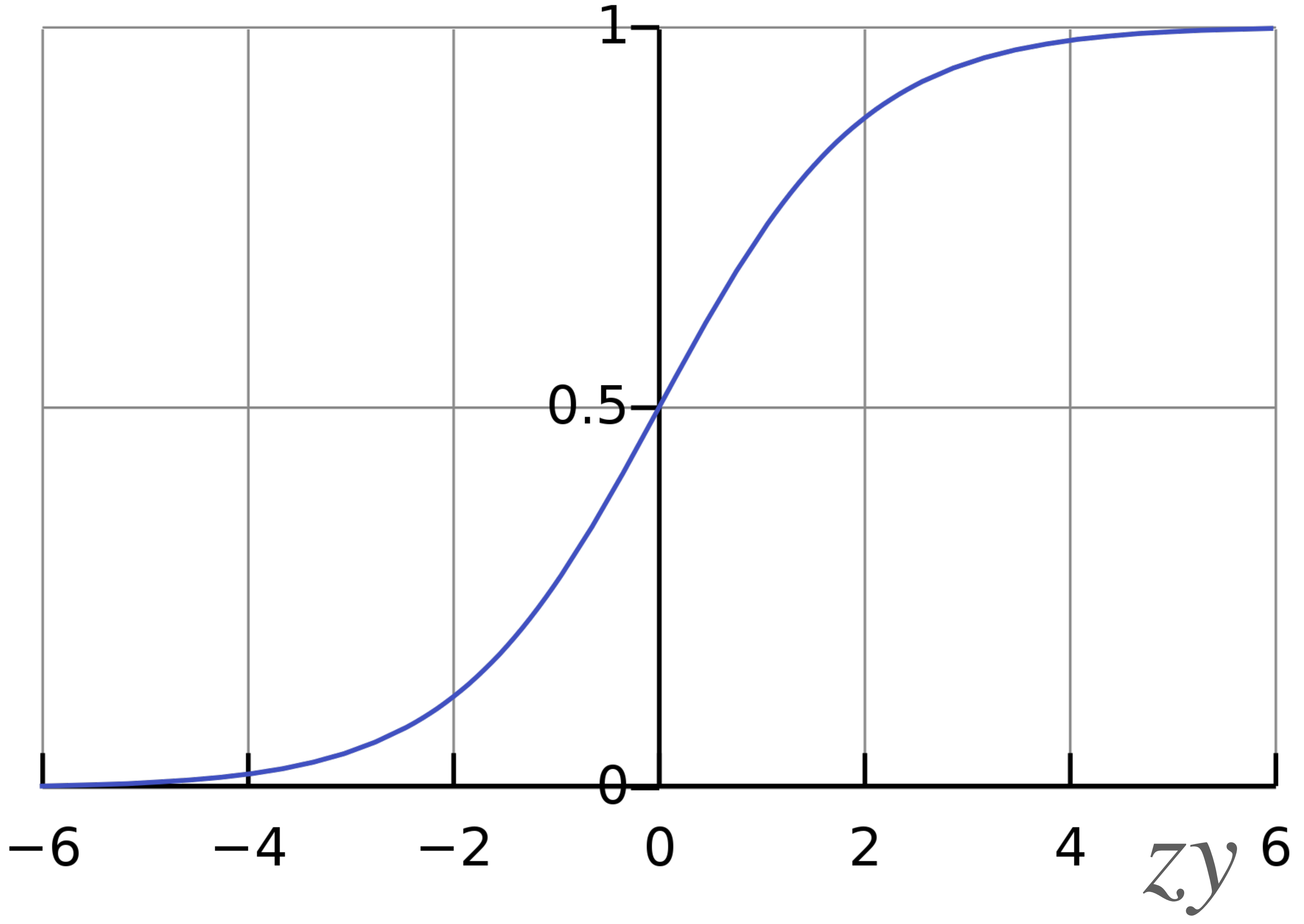
$$P(y \mid \overrightarrow{x}) = \frac{1}{1 + exp(-w^{\top}\frac{\overrightarrow{x}y}{Z})}$$

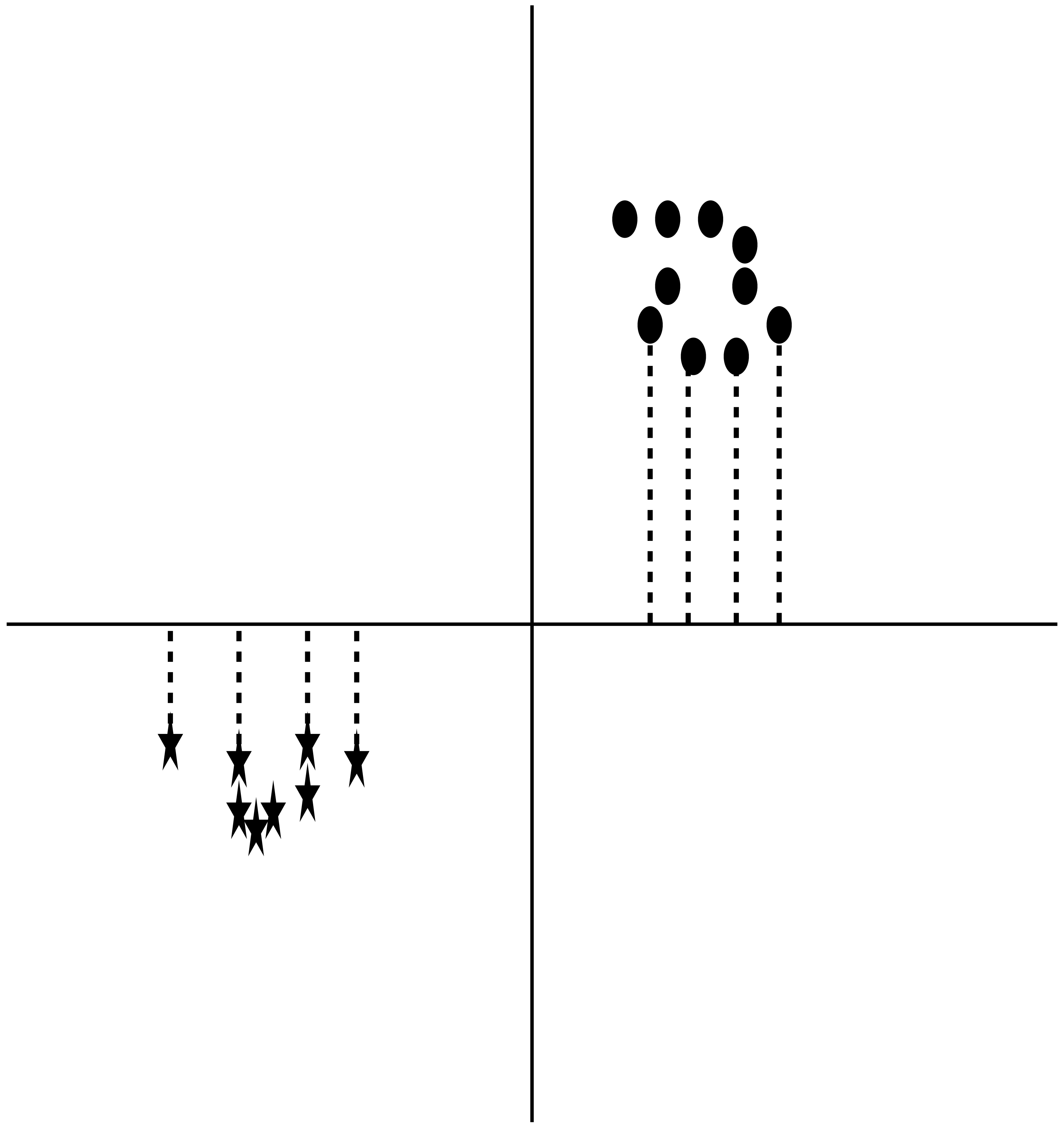
where, $y = \pm 1$



Generative v/s discriminative approach

$$P(y \mid \overrightarrow{x}) = \frac{1}{1 + exp(-w^{\top}\frac{\overrightarrow{x}y}{Z})}$$





MLE

Estimate w & b directly!

$$w, b = \arg\max_{w,b} \prod_{i=1}^{n} P_w(y_i \mid \overrightarrow{x}; w)$$

Subsuming b in w & taking logs

$$= \arg\max_{w} = \sum_{i=1}^{n} \log P_w(y_i \mid \overrightarrow{x}, w)$$

$$= \arg\max_{w} \sum_{i=1}^{n} \log \frac{1}{1 + exp(-yw^\top \overrightarrow{x})}$$

Logistic function

$$= \arg\max_{w} - \sum_{i=1}^{n} \log(1 + exp(-yw^\top \overrightarrow{x}))$$

$$= \arg\min_{w} \sum_{i=1}^{n} \log(1 + exp(-yw^\top \overrightarrow{x}))$$

Next week

Logistic regression

Cross entropy

Bayesian networks: formulation and inference