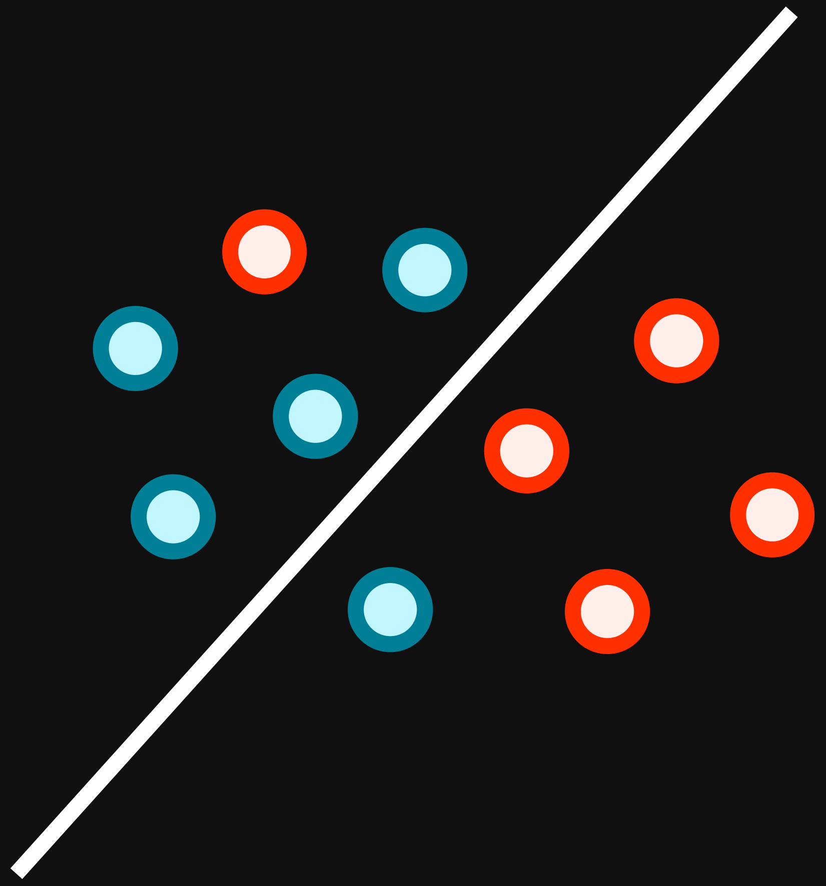
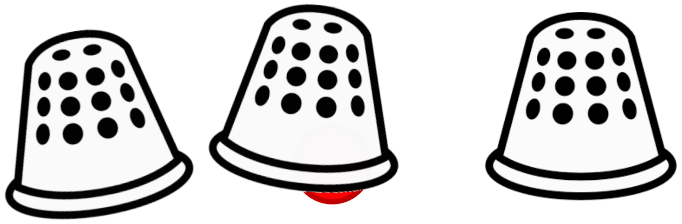
Basics of Machine Learning

Dmitry Ryabokon, github.com/dryabokon



Lesson 07
Bayesian Approach



Summary

- Approaches for ML
- Bayesian approach
- Classifying the independent binary features

Approaches for ML

Maximum Likelihood

$x_1, x_2, \dots, x_N$
$k_1, k_2, \dots, k_N$ features and states

$k \in \mathrm{K} = \{1, 2, \dots, K\}$ domain for *k*

$p(x|k) = f(x, \mu_k)$ The model is parametrized by $\mu_1, \mu_2, \dots, \mu_K$

input

$$(\mu_1, \mu_2, \dots, \mu_K)^* = \arg\max_{\mu_1, \mu_2, \dots, \mu_K} \prod_{i=1}^{N} p(x_i, k_i)$$

output

Approaches for ML

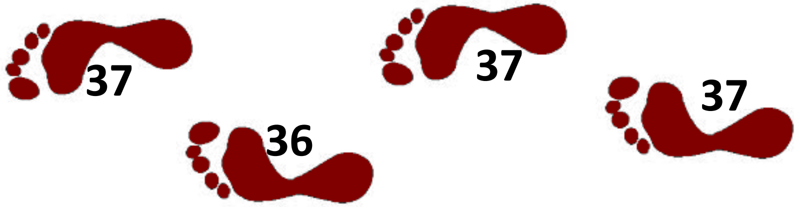
Maximum Likelihood

$$\arg\max_{\mu_1,\mu_2,\,\dots\,,\mu_K} \prod_{i=1}^{N} p(x_i, k_i) = \arg\max_{\mu_1,\mu_2,\,\dots\,,\mu_K} \left(\prod_{i=1}^{N} p(k_i) \cdot p(x_i | k_i) \right) = \arg\max_{\mu_1,\mu_2,\,\dots\,,\mu_K} \left(\prod_{i=1}^{N} p(x_i | k_i) \right) =$$

$$= \arg\max_{\mu_1,\mu_2,\,\dots\,,\mu_K} \left(\prod_{x \in X_1} p(x|1) \times \prod_{x \in X_2} p(x|2) \times \prod_{x \in X_K} p(x|K) \right)$$

$$\mu_k^{\,*} = \arg\max_{\mu} \prod_{x \in X_k} f(x, \mu)$$

Maximum Likelihood: example



$$p(x|k=1) = f(x,\mu_1) \cong \exp\left\{-(x-\mu_1)^2\right\}$$

$$p(x|k=2) = f(x,\mu_2) \cong \exp\left\{-(x-\mu_2)^2\right\}$$

$$(\mu_1,\mu_2) = \arg\max \prod_{i=1}^{7} p(x_i,k_i)$$

$$\mu_1 = \arg\max_{\mu} \quad \exp\left\{-(45-\mu)^2 - (46-\mu)^2 - (45-\mu)^2\right\}$$

Approaches for ML

Maximum Likelihood: example

$$\mu_1 = \arg\max_{\mu} \ \exp\left\{-(45-\mu)^2 - (46-\mu)^2 - (45-\mu)^2\right\}$$

$$\mu_1 = \arg\min_{\mu} \ \left\{(45-\mu)^2 + (46-\mu)^2 + (45-\mu)^2\right\}$$

$$\mu_1 = \frac{45+46+45}{3} = \langle x \rangle_{X_1}$$

Approaches for ML

Bias in training data

$x_1, x_2, \ldots, x_N$
$k_1, k_2, \ldots, k_N$ features and states

$k \in \mathrm{K} = \{1, 2, \ldots, K\}$ domain for *k*

$p(x|k) = f(x, \mu_k)$ The model is parametrized by $\mu_1, \mu_2, \ldots, \mu_K$

input

$$\mu_1^* = \arg\max_{\mu} \left\{ \min_{x \in X_1} p(x|k=1, \mu) \right\}$$

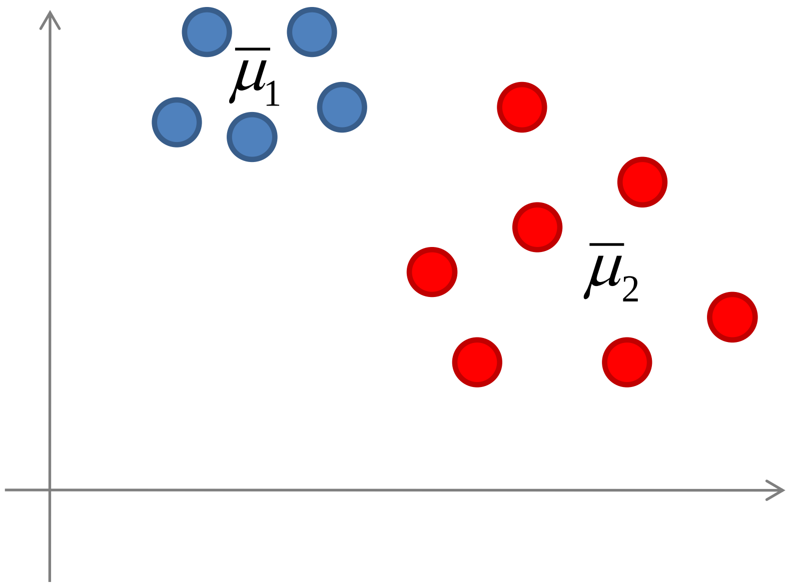
output

Approaches for ML

Bias in training data: example

$$p(x|k=1) = N(\bar{\mu}_1, 1)$$
$$p(x|k=2) = N(\bar{\mu}_2, 1)$$

$$\bar{\mu}_1^{\ *} = \arg\max_{\bar{\mu}} \left\{ \min_{x \in X_1} p(x|k=1, \bar{\mu}) \right\}$$



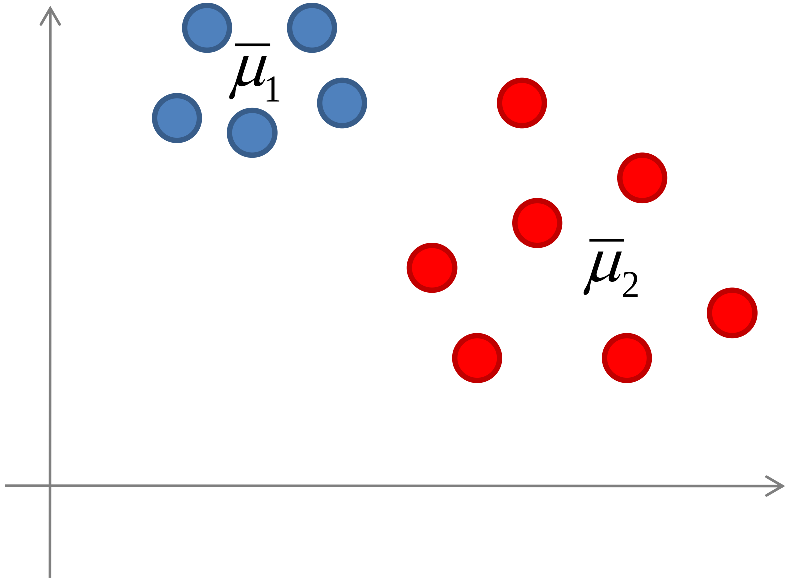
Approaches for ML

Bias in training data: example

$$p(x|k=1) = N(\bar{\mu}_1, 1)$$
$$p(x|k=2) = N(\bar{\mu}_2, 1)$$

$$\bar{\mu}_1^* = \arg\max_{\bar{\mu}} \left\{ \min_{x \in X_1} p(x|k=1, \bar{\mu}) \right\}$$

center of the circle that
holds all the points from X1



Approaches for ML

ERM: Empirical risk minimization

$x_1, x_2, \ldots, x_N$ features and
$k_1, k_2, \ldots, k_N$ states

$k \in \mathrm{K} = \{1, 2, \ldots, K\}$ domain for *k*

$W(k, k')$ penalty function

$q(x, \overline{\mu}) \in K$ decision strategy is parametrized by μ

ERM: Empirical risk minimization

$q(x, \overline{\mu}) \in K$ decision strategy is parametrized by μ

$$Risk\left(q(\overline{\mu})\right) = \sum_{x \in X} \sum_{k \in K} p(x,k) \cdot W\left(q(x,\overline{\mu}),\ k\right)$$

$$\cong \sum_{i=1}^{N} p(x_i, k_i) \cdot W\left(q(x_i,\overline{\mu}),\ k_i\right) = \sum_{i=1}^{N} W\left(q(x_i,\overline{\mu}),\ k_i\right)$$

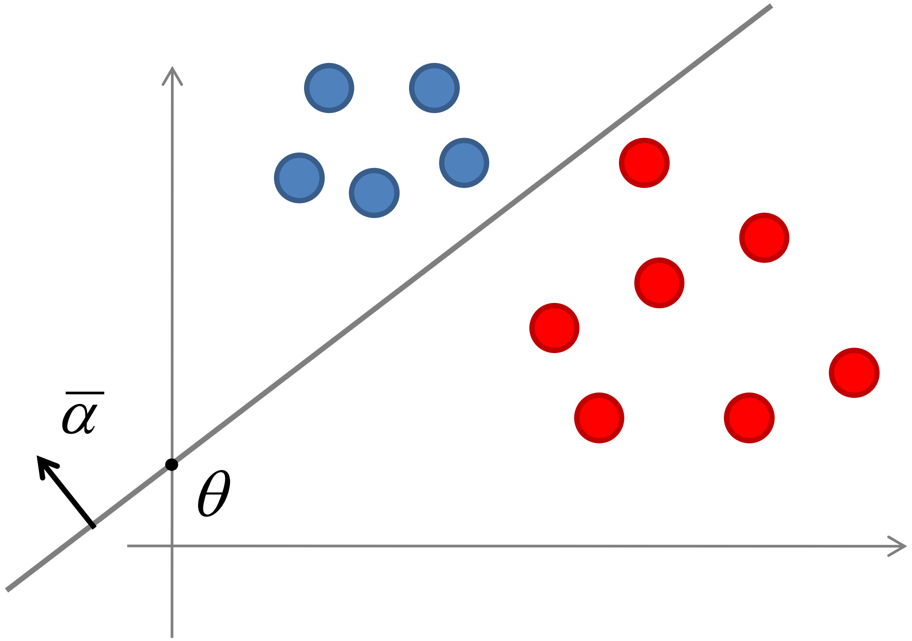
$$\mu^* = \arg\min_{\mu}\ Risk\ q(\mu)$$

ERM: Empirical risk minimization

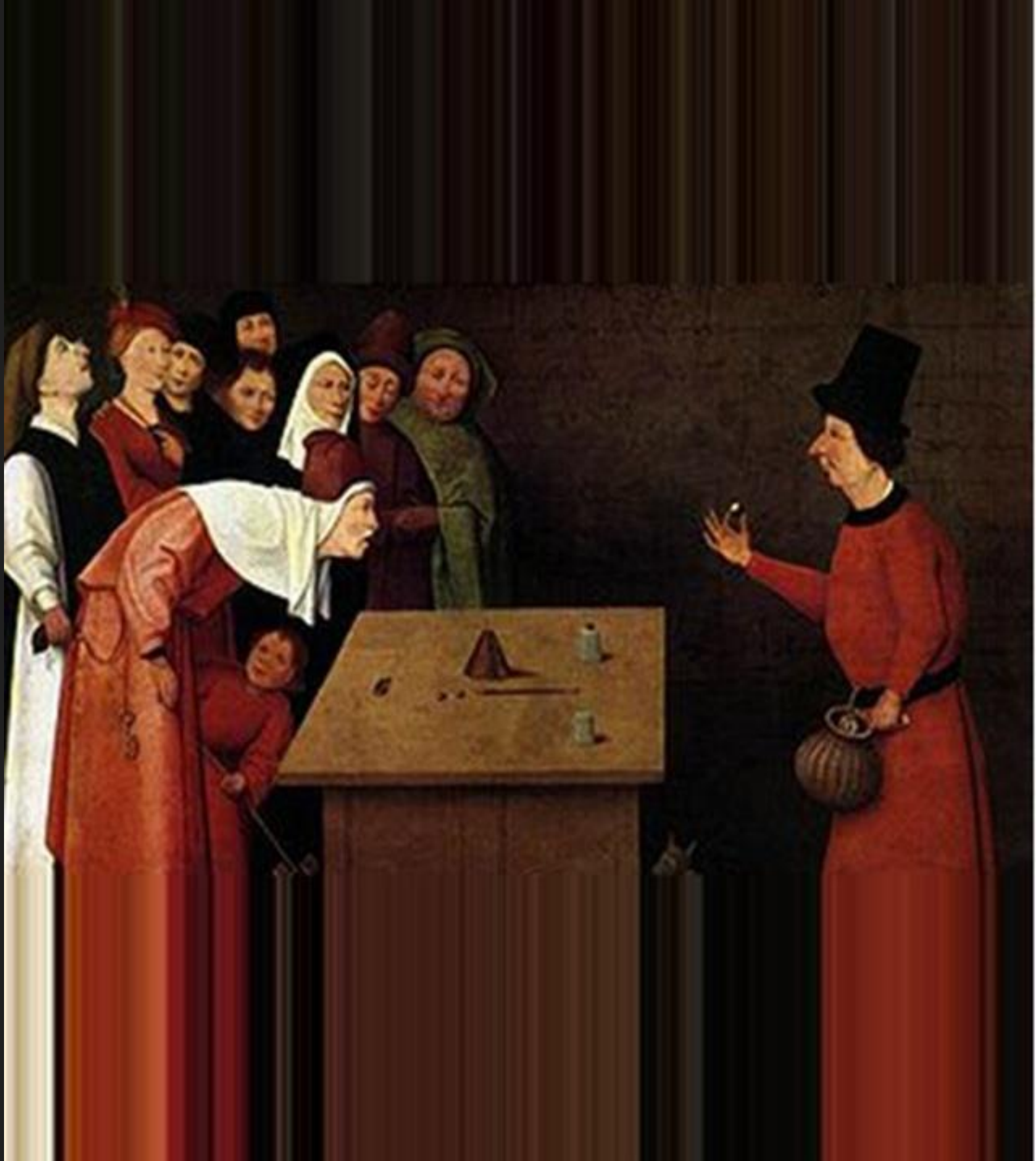
$$q(\bar{x},\bar{\alpha},\theta) = \begin{cases} -1 & when\ (\bar{x},\bar{\alpha}) < \theta \\ +1 & when\ (\bar{x},\bar{\alpha}) > \theta \end{cases}$$

$$W(k,k') = \begin{cases} 0 & k = k' \\ 1 & k \neq k' \end{cases}$$

$$Risk(q(\overline{\alpha},\theta)) =$$ Number of errors



Bayesian approach



Bayesian approach

Definitions

$x \in X$ feature

$k \in \mathrm{K}$ state (hidden)

$p(x,k)$ joint probability

$W : \mathrm{K} \times \mathrm{K} \to R$ penalty function

$q : X \to \mathrm{K}$ decision strategy

$$Risk(q) = \sum_{k \in \mathrm{K}} \sum_{x \in X} p(x,k) \cdot W(k, q(x))$$

Bayesian approach

The problem

$x \in X$ feature

$k \in K$ state (hidden)

$p(x,k)$ joint probability

$W : K \times K \to R$ penalty function

input

$q : X \to K$ decision strategy

$$q^*(x) = \arg\min_{q(x)} Risk\left(q(x)\right) = \arg\min_{q(x)} \sum_{k \in K} \sum_{x \in X} p(x,k) \cdot W\left(k, q(x)\right)$$

output

Bayesian approach

Some basics

$$p(x,k) = p(k)\cdot p(x|k) = p(x)\cdot p(k|x)$$ joint probability

$$p(x) = \sum_k p(k)\cdot p(x|k)$$ law of total probability

$$p(k|x) = \frac{p(k)\cdot p(x|k)}{\sum_{k'} p(k')\cdot p(x|k')}$$ Bayes' theorem

---

p(k)	prior probability
p(k|x)	posterior probability
p(x|k)	conditional probability of *x*, assuming *k*
p(x)	probability of *x*

Bayesian approach

Example

$X = \{8, 9, 10, 11, 12\}$ time

К = {1-revision, 2-free_ride}

$$p(k=1 | x=8) = 10\%$$
$$p(k=2 | x=8) = 90\%$$

$W(k=1, k'=1) = 5$ $W(k=2, k'=1) = 100$

$W(k=1, k'=2) = 5$ $W(k=2, k'=2) = 0$

Bayesian approach

Flexibility to not make a decision

$$W(k, k') = \begin{cases} 0 & when\ k = k' \\ 1 & when\ k \neq k' \\ \varepsilon & when\ k = reject \end{cases}$$

$$Risk(reject) = \sum_{k \in K} p(k|x) \cdot W(k, reject) =$$

$$= \sum_{k \in K} p(k|x) \cdot \varepsilon = \varepsilon$$

Bayesian approach

Flexibility to not make a decision

$$\varepsilon = 0$$

if $\min_k Risk(k) < \varepsilon$

then $\cancel{k^* = \arg\min_k \sum_{k' \in K} p(k|x) \cdot W(k,k')}$

else reject

$$\varepsilon = \infty$$

if $\min_k Risk(k) < \varepsilon$

then $k^* = \arg\min_k \sum_{k' \in K} p(k|x) \cdot W(k,k')$

else ~~reject~~

Bayesian approach

Decision strategies for different penalty functions

$$w(k,k') = |k - k'|$$

median

$$w(k,k') = (k - k')^2$$

average

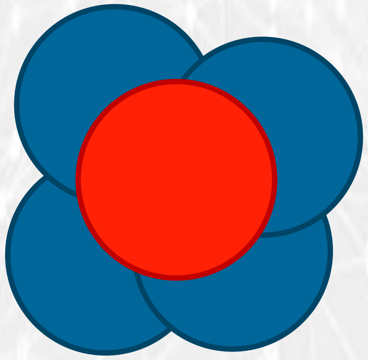
$$w(k,k') = \begin{cases} 0 & k = k' \\ 1 & k \neq k' \end{cases}$$

the most probable sample

$$w(k,k') = \begin{cases} 0 & |k - k'| < \Delta \\ 1 & |k - k'| \geq \Delta \end{cases}$$

center of the most probable section Δ

Classifying the independent binary features



Classifying the independent binary features

The problem

$\bar{x} = (x_1, x_2, \ldots, x_N)$ feature – a set of independent binary values

$$p(\bar{x}|k) = p(x_1|k) \cdot p(x_2|k) \cdot \ldots \cdot p(x_N|k)$$

$k \in \mathrm{K} = \{1, 2\}$ Few states are possible

$$\frac{p(\bar{x}|k=1)}{p(\bar{x}|k=2)} \mathop{\gtrless}_{2}^{1} \theta$$

Decision strategy

Classifying the independent binary features

The problem

$$\log \frac{p(\bar{x}|k=1)}{p(\bar{x}|k=2)} = \log \frac{p(x_1|k=1)\cdot\ldots\cdot p(x_n|k=1)}{p(x_1|k=2)\cdot\ldots\cdot p(x_n|k=2)} = \log \frac{p(x_1|k=1)}{p(x_1|k=2)} + \ldots + \log \frac{p(x_n|k=1)}{p(x_n|k=2)} =$$

$$x_1 \cdot \log \frac{p(x_1=1|k=1)\cdot p(x_1=0|k=2)}{p(x_1=1|k=2)\cdot p(x_1=0|k=1)} + \log \frac{p(x_1=0|k=1)}{p(x_1=0|k=2)} +$$

$$\vdots$$

$$x_N \cdot \log \frac{p(x_n=1|k=1)\cdot p(x_n=0|k=2)}{p(x_n=1|k=2)\cdot p(x_n=0|k=1)} + \log \frac{p(x_n=0|k=1)}{p(x_n=0|k=2)}$$

Decision rule

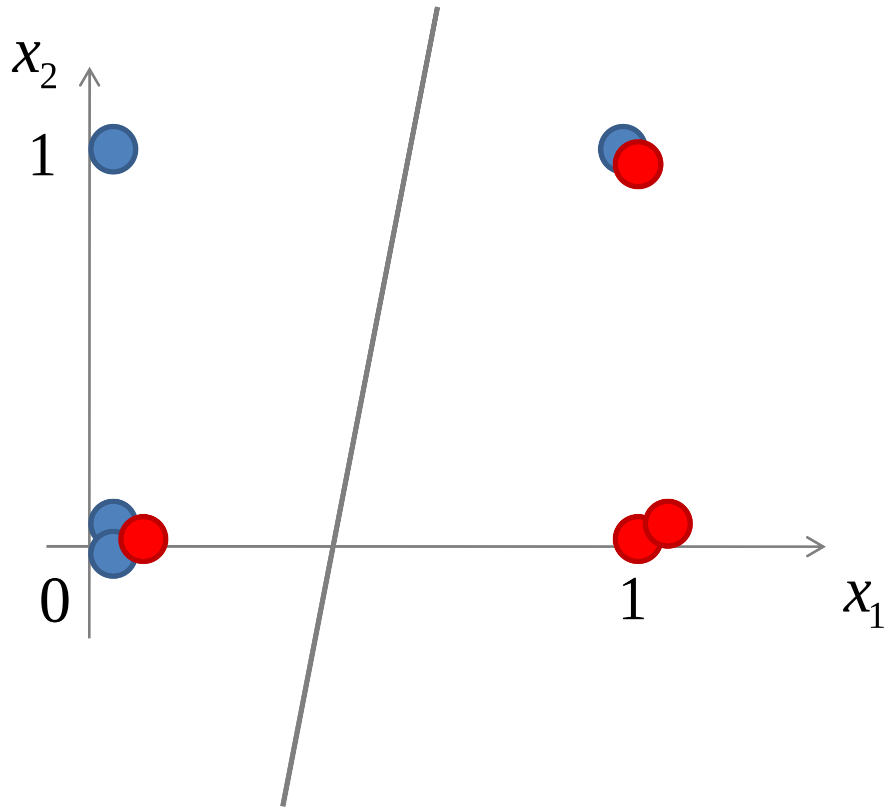
$$\sum_{i=1}^{N} x_i \cdot \log \frac{p(x_i = 1|k=1)\cdot p(x_i = 0|k=2)}{p(x_i = 1|k=2)\cdot p(x_i = 0|k=1)} + \log \frac{p(x_i = 0|k=1)}{p(x_i = 0|k=2)} \;\underset{2}{\overset{1}{\gtreqless}}\; \log \theta$$

$$\boxed{\sum_{i=1}^{N} x_i \cdot a_i \;\underset{2}{\overset{1}{\gtreqless}}\; \theta'}$$

Classifying the independent binary features

Example

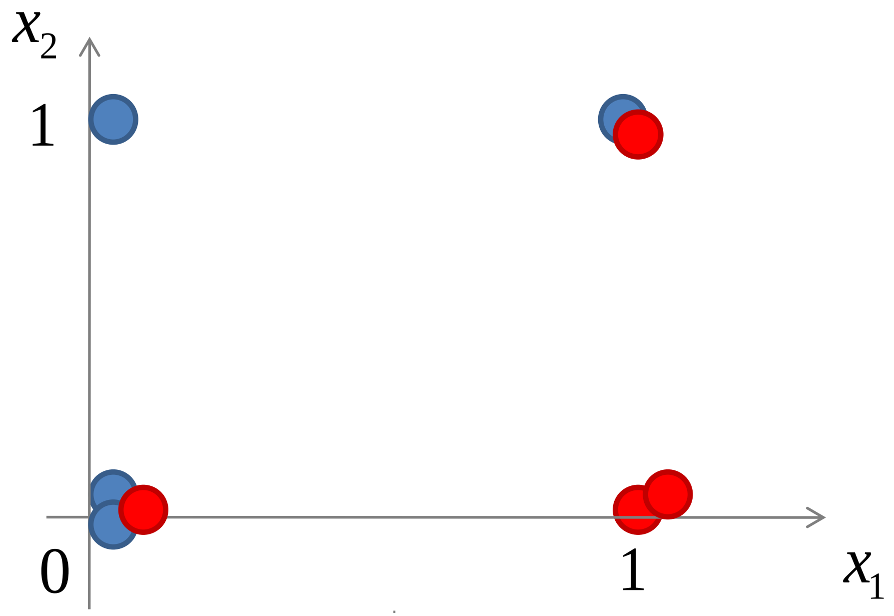
$$\sum_{i=1}^{N} x_i \cdot a_i \;\begin{array}{c} 1 \\ \geq \\ \leq \\ 2 \end{array}\; \theta'$$



Classifying the independent binary features

Example

$p(x_2 \mid {\color{red}\bullet})$	$p(x_2 \mid {\color{blue}\bullet})$
1/4	**2/4**
3/4	**2/4**

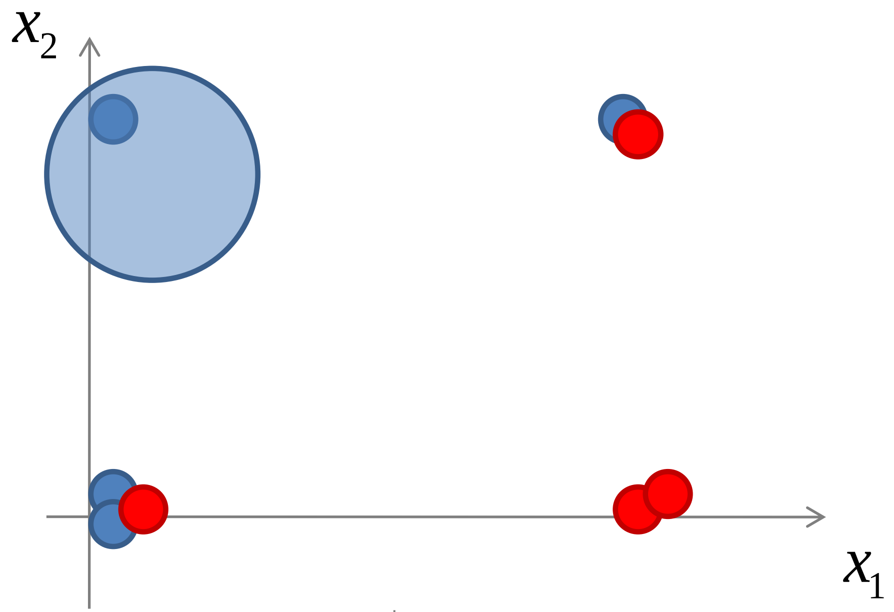


1/4	**3/4**	$p(x_1 \mid {\color{red}\bullet})$
3/4	**1/4**	$p(x_1 \mid {\color{blue}\bullet})$

Classifying the independent binary features

Example

$p(x_2 \mid \bullet)$	$p(x_2 \mid \bullet)$
$\boxed{1/4}$	$\boxed{2/4}$
3/4	2/4



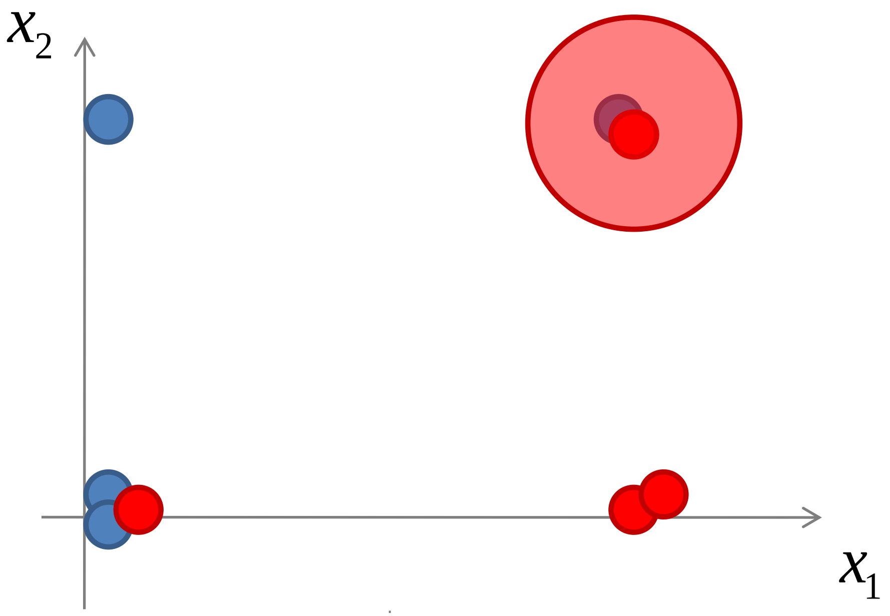
$\boxed{1/4}$	3/4	$p(x_1 \mid \bullet)$
$\boxed{3/4}$	1/4	$p(x_1 \mid \bullet)$

$$(1/4)\text{x}(1/4) < \boxed{(2/4)\text{x}(3/4)}$$

Classifying the independent binary features

Example

$p(x_2 \mid \bullet)$	$p(x_2 \mid \bullet)$
$\boxed{1/4}$	$\boxed{2/4}$
3/4	2/4



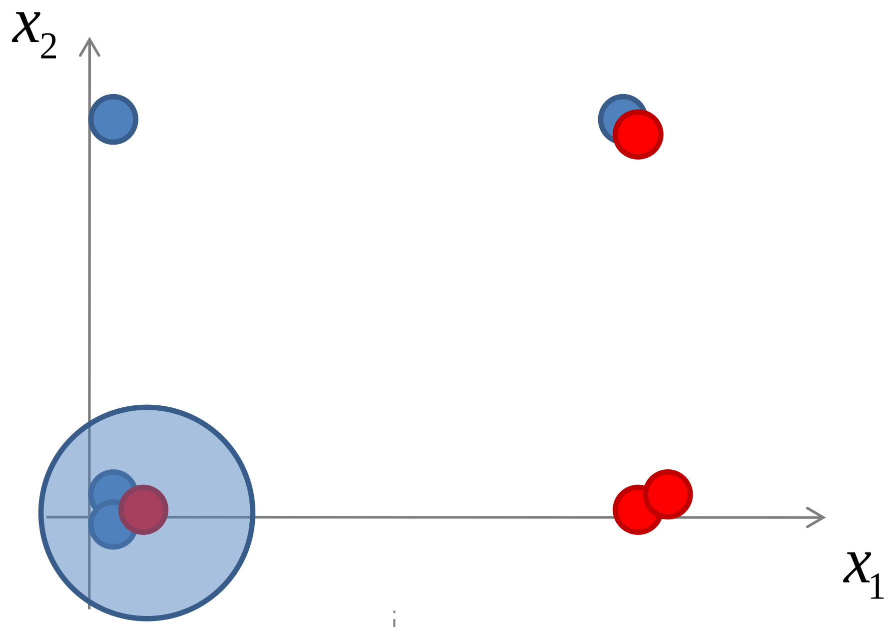
1/4	$\boxed{3/4}$	$p(x_1 \mid \bullet)$
3/4	$\boxed{1/4}$	$p(x_1 \mid \bullet)$

$\boxed{(1/4)\times(3/4)}$ > $(2/4)\times(1/4)$

Classifying the independent binary features

Example

$p(x_2 \mid ●)$	$p(x_2 \mid ●)$
1/4	2/4
$\boxed{3/4}$	$\boxed{2/4}$



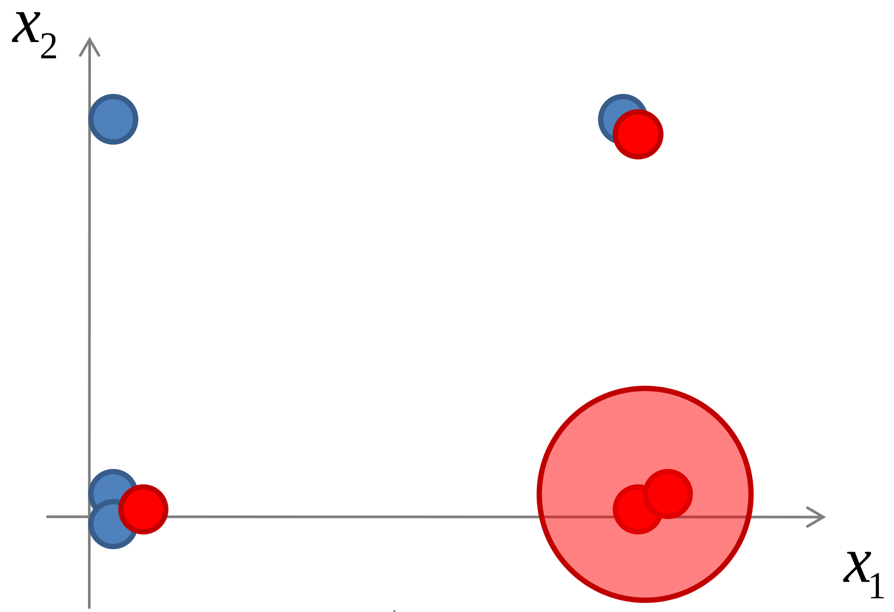
$\boxed{1/4}$	3/4	$p(x_1 \mid ●)$
$\boxed{3/4}$	1/4	$p(x_1 \mid ●)$

(3/4)x(1/4) < $\boxed{(2/4)x(3/4)}$

Classifying the independent binary features

Example

$p(x_2 \mid \bullet)$	$p(x_2 \mid \bullet)$
1/4	**2/4**
$\boxed{\textbf{3/4}}$	$\boxed{\textbf{2/4}}$

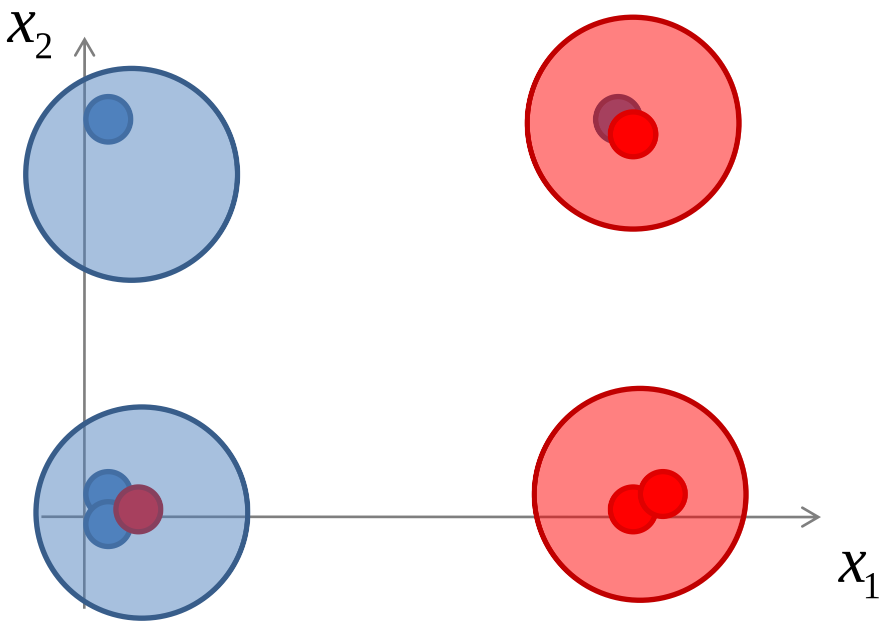


1/4	$\boxed{\textbf{3/4}}$	$p(x_1 \mid \bullet)$
3/4	$\boxed{\textbf{1/4}}$	$p(x_1 \mid \bullet)$

$\boxed{(3/4) \times (3/4)}$ **>** $(2/4) \times (1/4)$

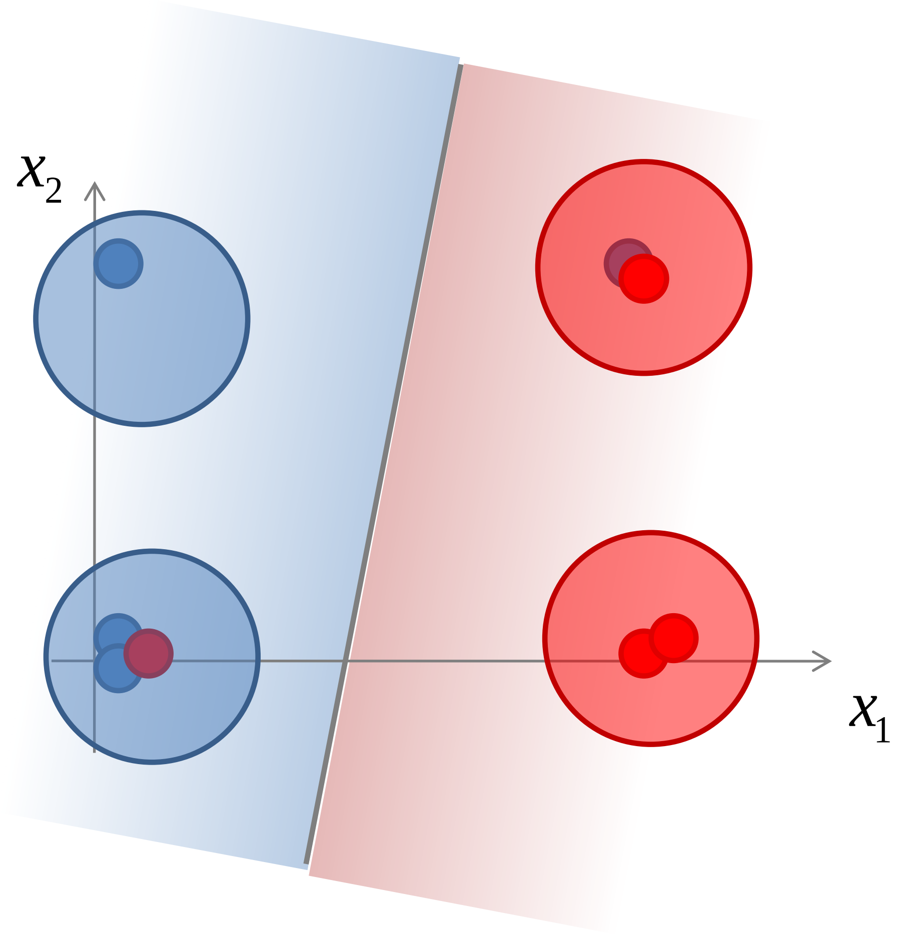
Classifying the independent binary features

Example



Classifying the independent binary features

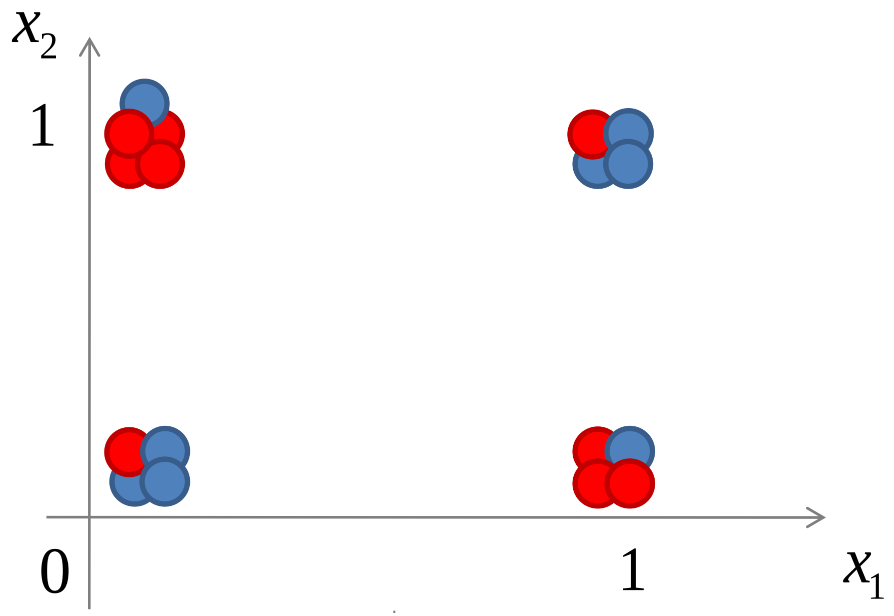
Example



Classifying the independent binary features

Another example

$p(x_2 \mid ●)$	$p(x_2 \mid ●)$
5/9	**4/8**
4/9	**4/8**

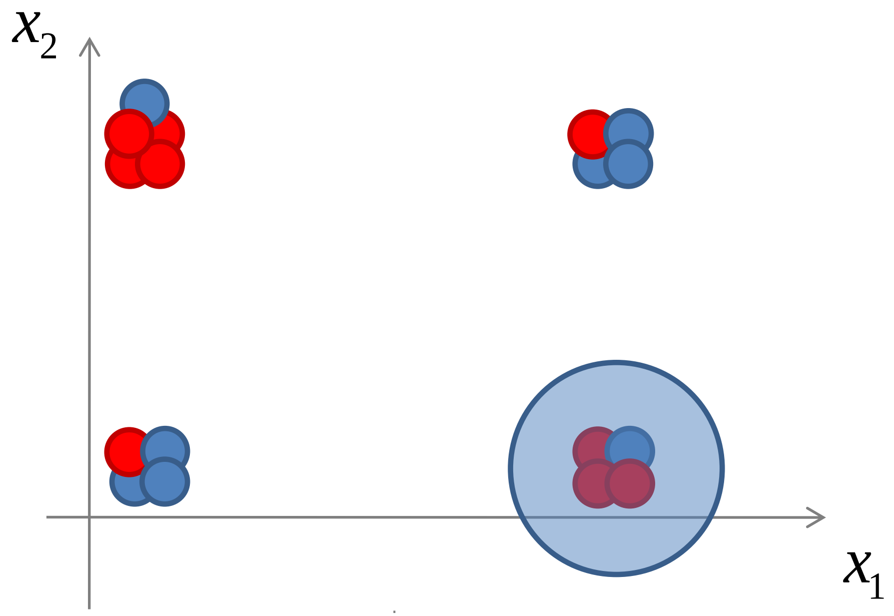


5/9	**4/8**	$p(x_1 \mid ●)$
4/8	**4/8**	$p(x_1 \mid ●)$

Classifying the independent binary features

Another example

$p(x_2 | \bullet)$ $p(x_2 | \bullet)$



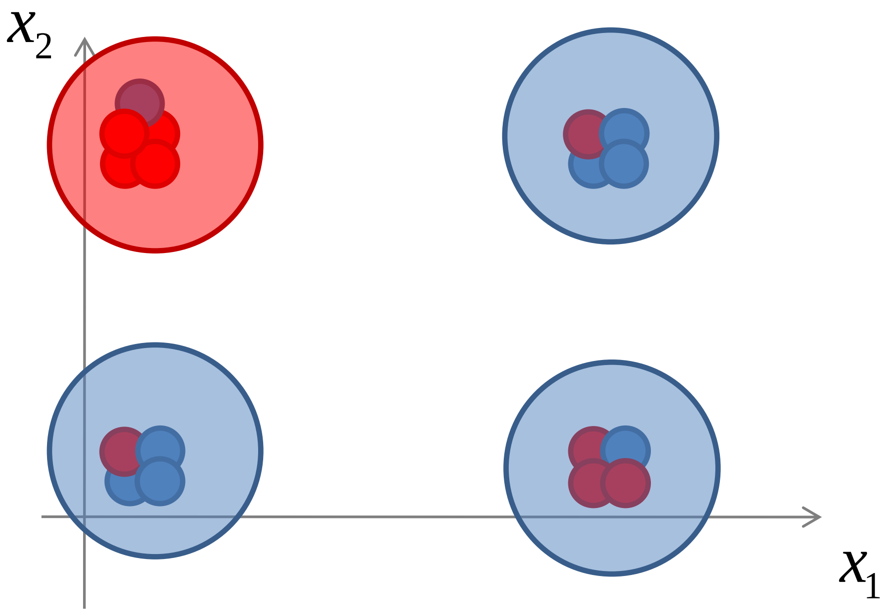
$p(x_1 | \bullet)$

$p(x_1 | \bullet)$

(4/9)x(4/9) < $\boxed{(4/8)x(4/8)}$

Classifying the independent binary features

Another example



Classifying the independent binary features

Another example