Statistics and Data Science for Engineers
E178 / ME276DS

Classification,
Naïve Bayes

© 2022 Gabriel Gomes

Recall

prediction function. $\theta \in \mathbb{R}^P$

- The prediction problem: $x \in \mathbb{R}^D \longrightarrow \boxed{h(x;\theta)} \xrightarrow{\hat{y}}$

$\hat{y} = h(x;\theta)$.

- Supervised learning: Given $\mathcal{D} = \{(x_i, y_i)\}_N$

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^P}{\text{argmin}} \sum_{i=1}^{N} L(y_i, h(x_i;\theta))$$

- Regression: $y \in \mathbb{R}$

$$L(y, \hat{y}) = L_2(y, \hat{y}) = (y - \hat{y})^2$$

- Linear regression: $h(x, \theta_0, \underline{\theta}_1) = \theta_0 + x^T \underline{\theta}_1 = \theta_0 + \theta_1 x^1 + \dots + \theta_D x^D$.

machine learning.

convex optimization.

$\Downarrow$

explicit solution to linear regression.

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$ ... explicit ... for any D.

Assumption "true system"

$\hat{\theta}$ linear functions of ε (noise/uncertainties). σ^2

$\longrightarrow$ statistical behavior of $\hat{\theta}$.

(unbiased and known variance) $\hat{\theta}_0, \hat{\theta}_1, \hat{y}$

Var

statistics.

Classification

- Sample space: $y \in \{c_1, c_2, \dots, c_K\}$ .... classes or labels.
- Loss function: ??? $(\text{cat} - \text{dog})^2$ K... # classes
- Prediction model:
  - Hard classifiers: $\bar{h}(x;\theta)$ returns a *class* $\hat{y} \in \{c_1, c_2, \dots, c_K\}$
  - Soft classifiers: $\tilde{h}(x;\theta)$ returns a *distribution* $P(\hat{y} \mid X = x)$ over classes.

{cat, no cat}.

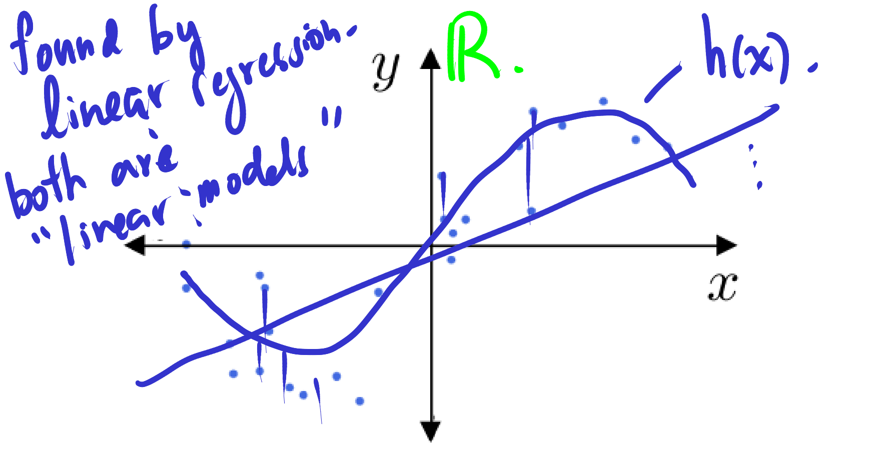
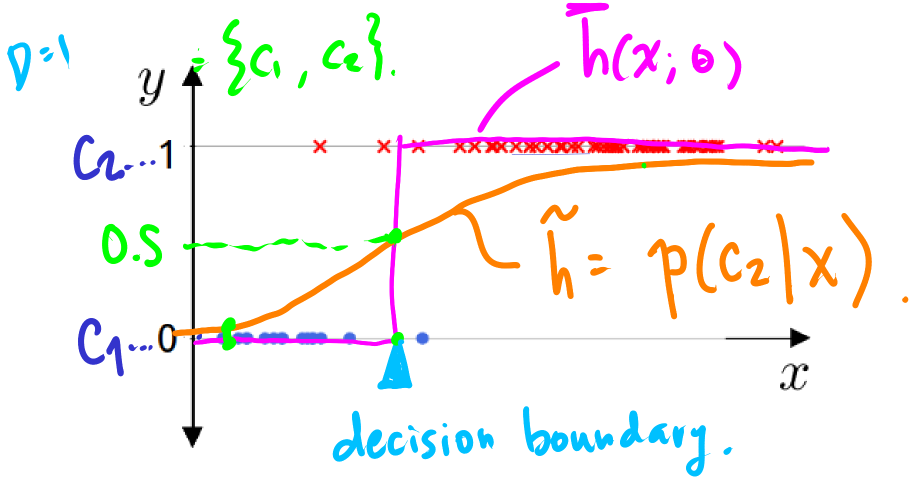
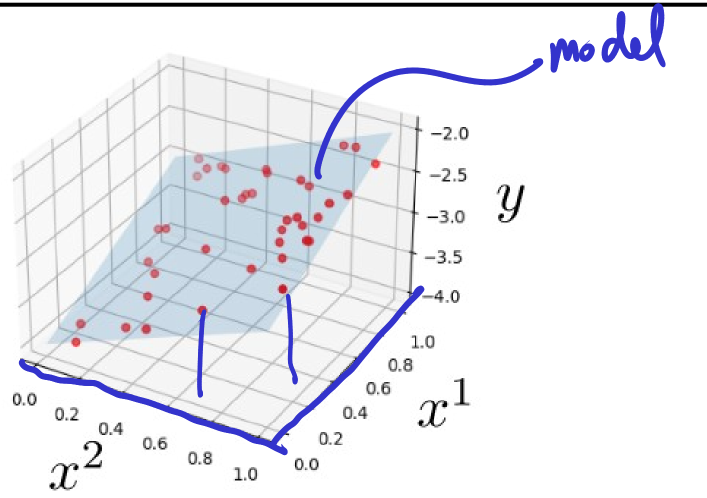
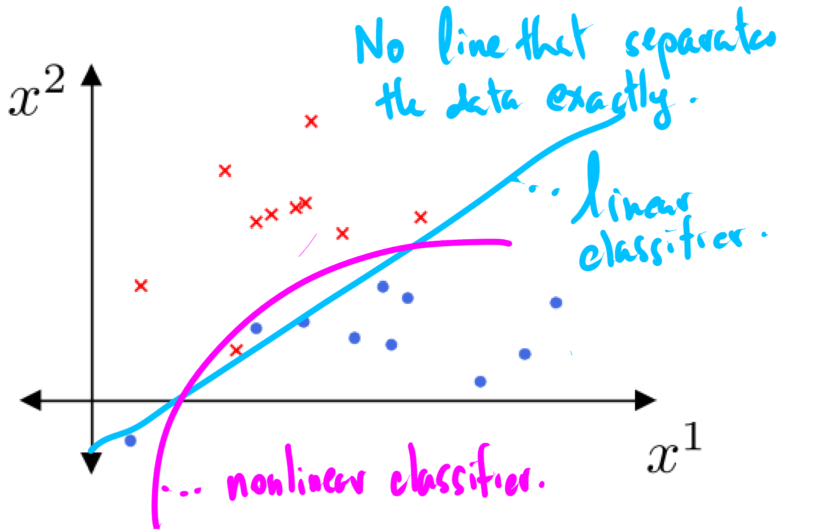
{cat, dog, person, sign, chair}.

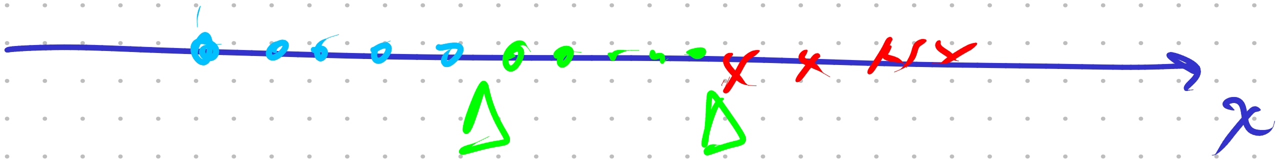
{benign, malignant}.

$$x \longrightarrow \boxed{\bar{h}} \longrightarrow \text{`cat'}$$

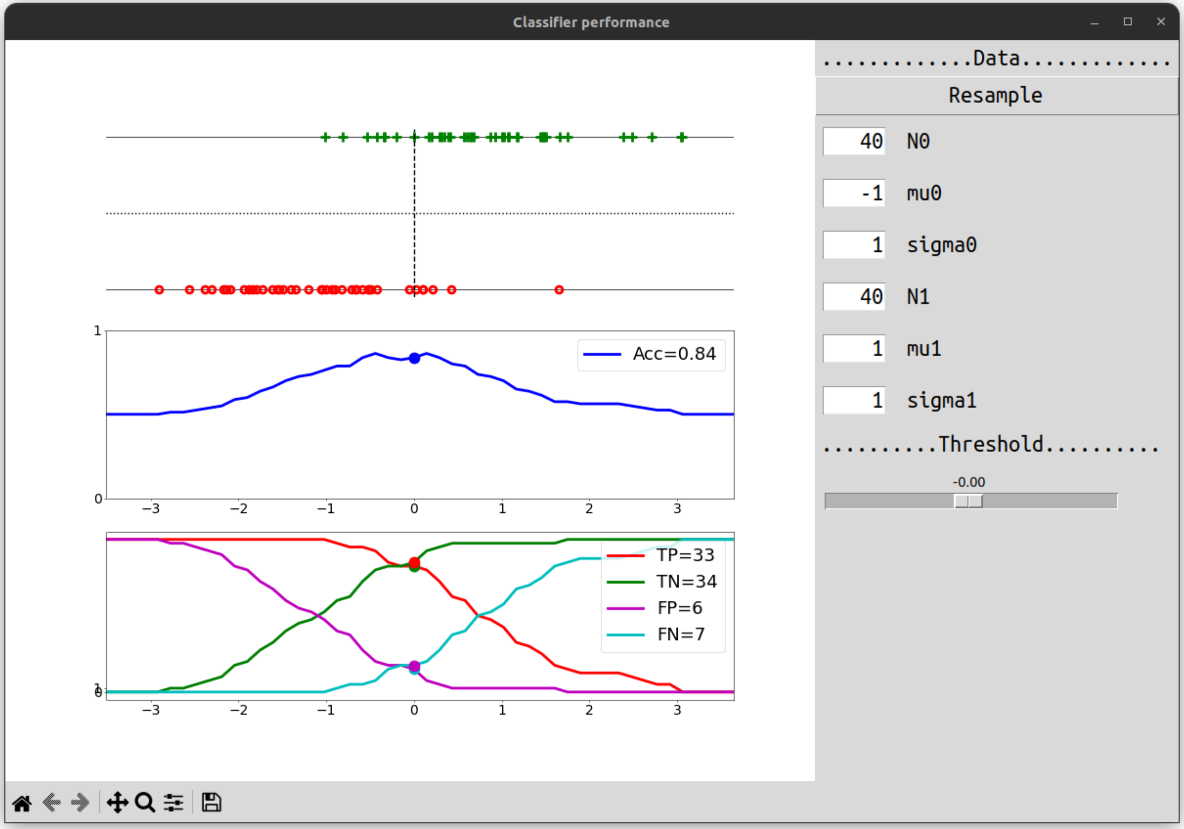
$$x \longrightarrow \boxed{\tilde{h}} \longrightarrow$$

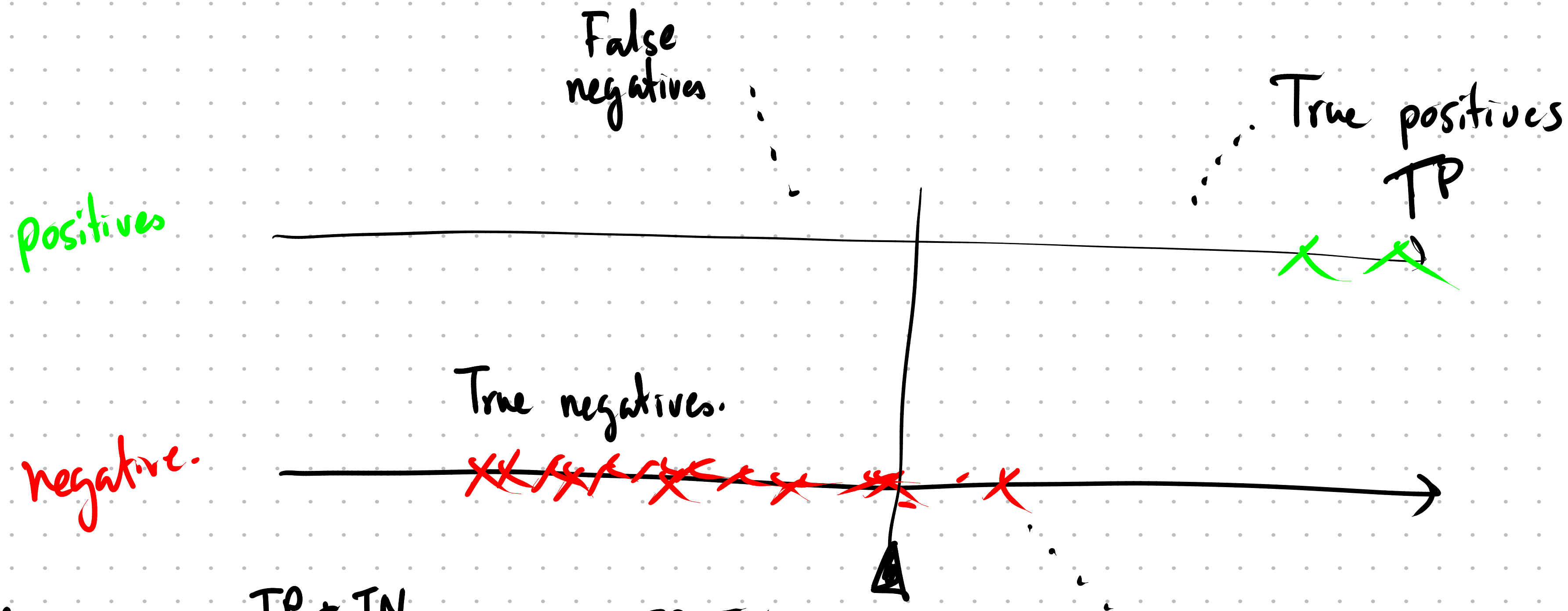
$$\bar{h}(x) = \arg\max_{c} \tilde{h}(x).$$

	Regression	K=2 Two-class classification
$D = 1$		
$D = 2$		


x

Demo: 1D, 2-class classification





$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{N}.$$

$h(x) = \text{negativ}$ $\qquad$ $h(x) = \text{positive}$

Accuracy is good when the problem is "symmetric"

- I have equal preference for FP and FN.
- The number of data points in each category is approx the same. (balanced dataset).

If "non-symmetric" then:

- precision, recall, balanced accuracy, TPR, TNR
- ROC curve
- AUC.

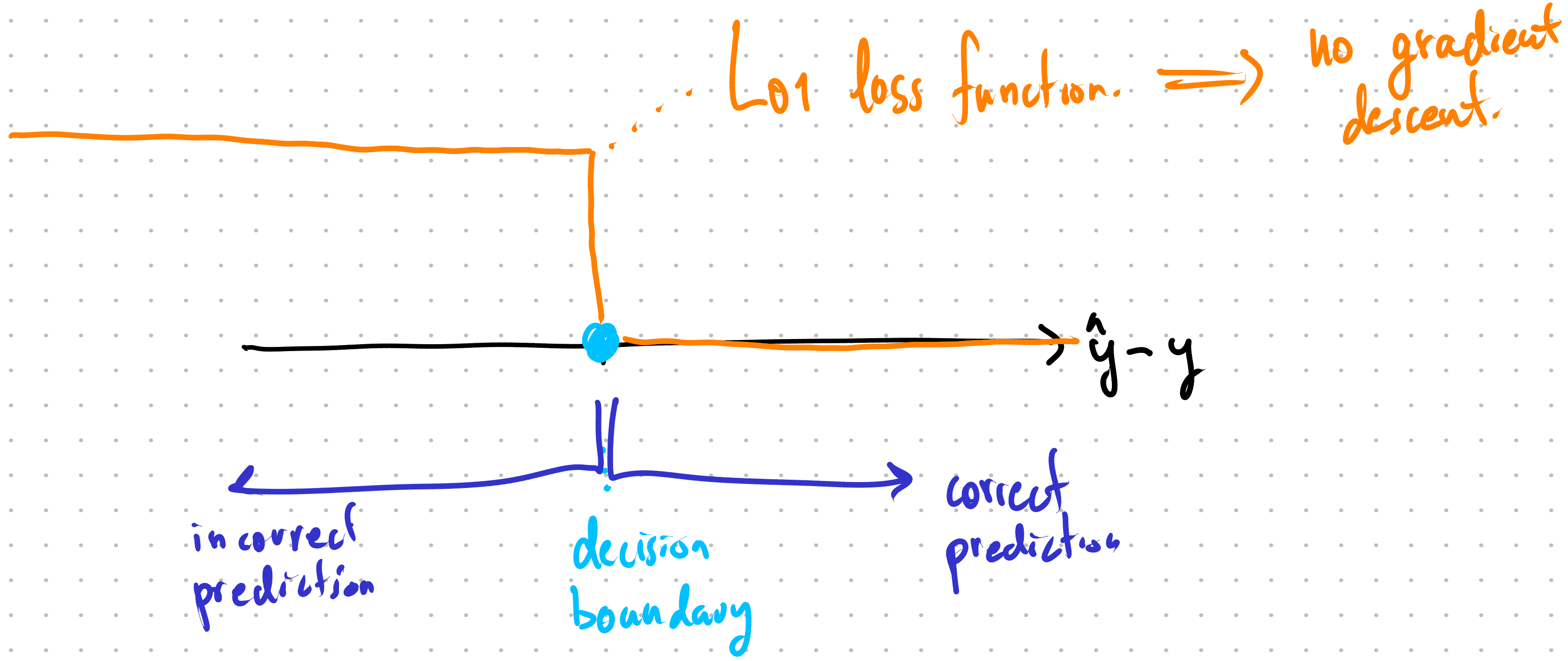
Accuracy as a loss function

$$L_{01}(y,\hat{y}) = \begin{cases} 1 & y \neq \hat{y} \\ 0 & y = \hat{y} \end{cases}$$

$$x \in \mathbb{R}^D \longrightarrow \boxed{\bar{h}(x;\theta)} \longrightarrow \hat{y} \in \{c_1, \dots, c_K\}$$

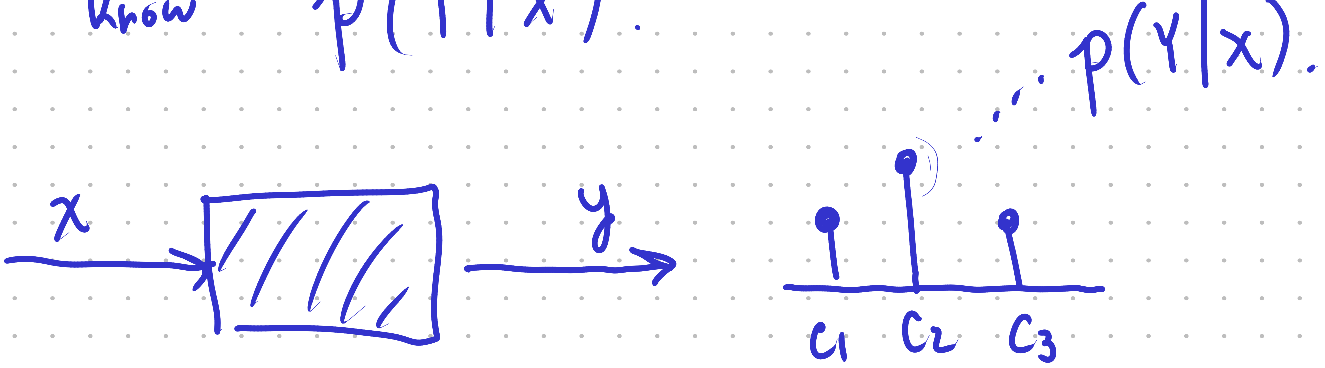
Accuracy is maximized by solving.

$$\text{minimize} \sum_{i=1}^{N} L_{01}\left(y, h(x;\theta)\right).$$



L01 loss function. ⟹ no gradient descent.
$\hat{y}-y$
incorrect prediction
decision boundary
correct prediction

- Cannot solve optimization problem directly with L_{01}
- Assume that we know $p(Y|x)$.



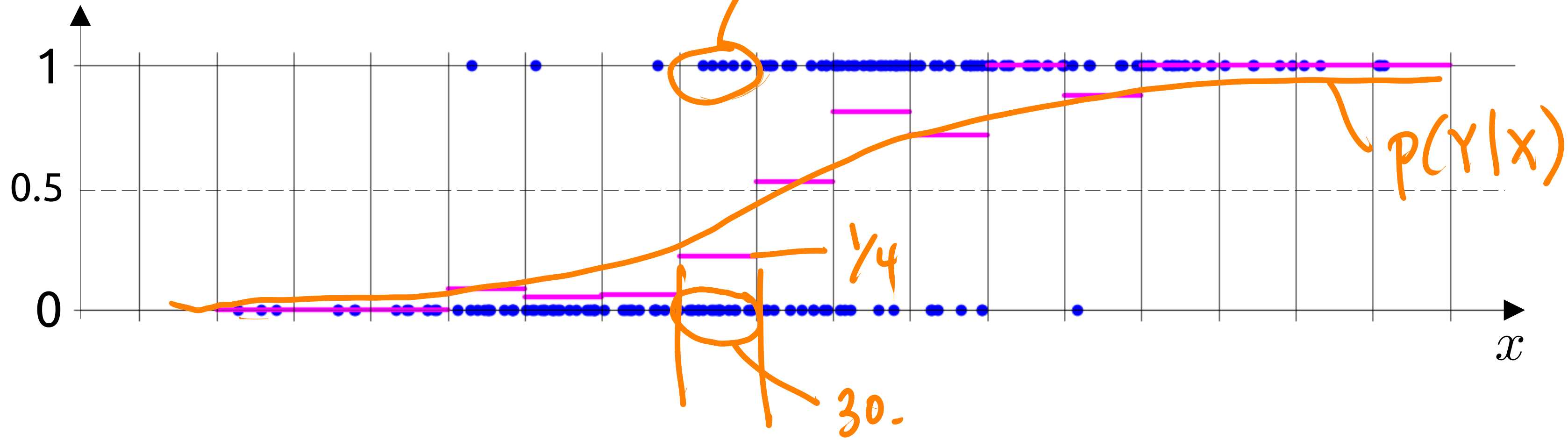
$$\hat{y} = \arg\max_{c} p(Y|X=x) \implies \text{minimize } E[L_{01}]$$

$$\implies \text{maximize accuracy.}$$

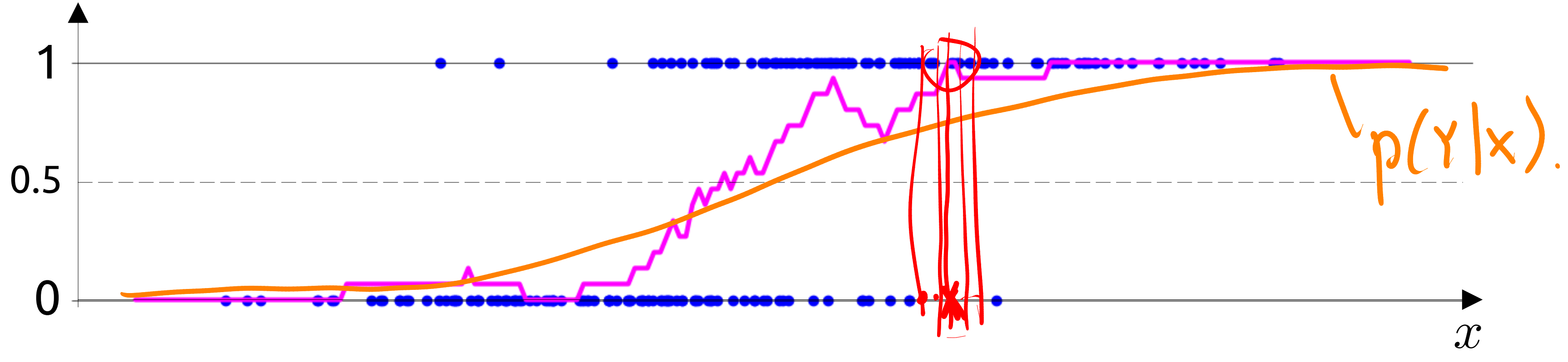
- Approximate $p(Y|X=x)$.

parametric.

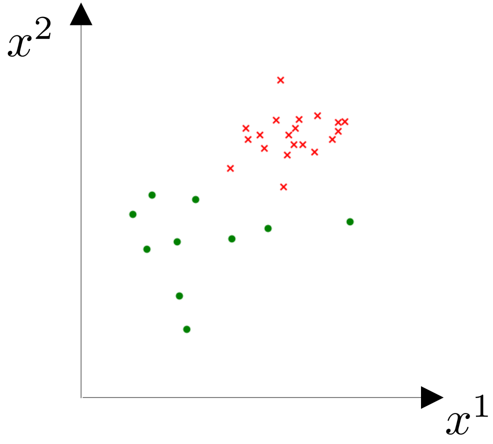
K-bins

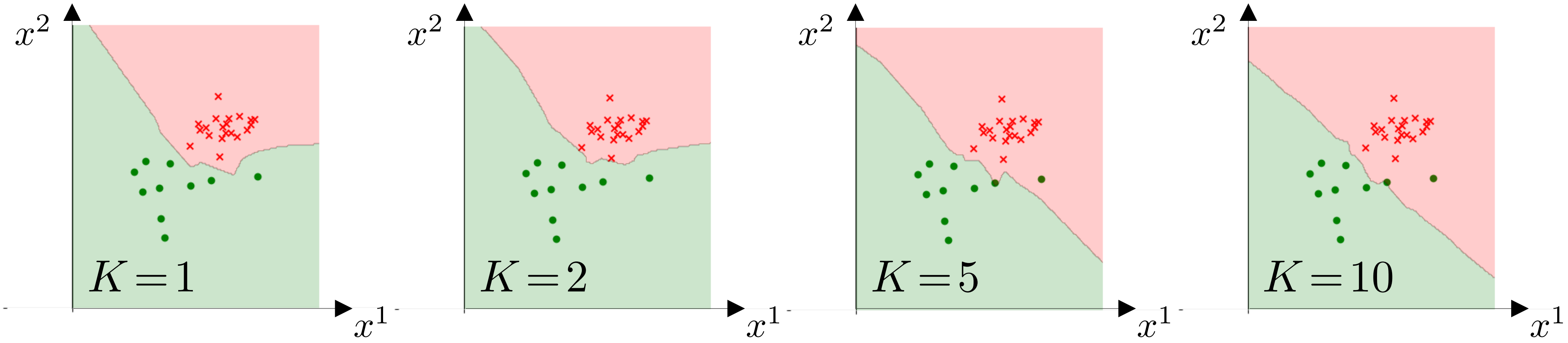


KNN



Example: Classifying two-class / 2D data with KNN





Applying Bayes' rule

$$\hat{y} = \underset{c \in \{c_1, \dots, c_K\}}{\text{argmax}} \; P(Y = c \mid X = x) \quad \text{I}$$

Bayes' rule.

$$= \underset{c \in \{c_1, \dots, c_K\}}{\text{argmax}} \; \frac{p(X = x \mid Y = c) P(Y = c)}{p(X = x)}$$

$$= \underset{c \in \{c_1, \dots, c_K\}}{\text{argmax}} \; p(X = x \mid Y = c) P(Y = c) \quad \text{II}$$

posterior belief.

likelihood of the input x given the class.

prior belief.

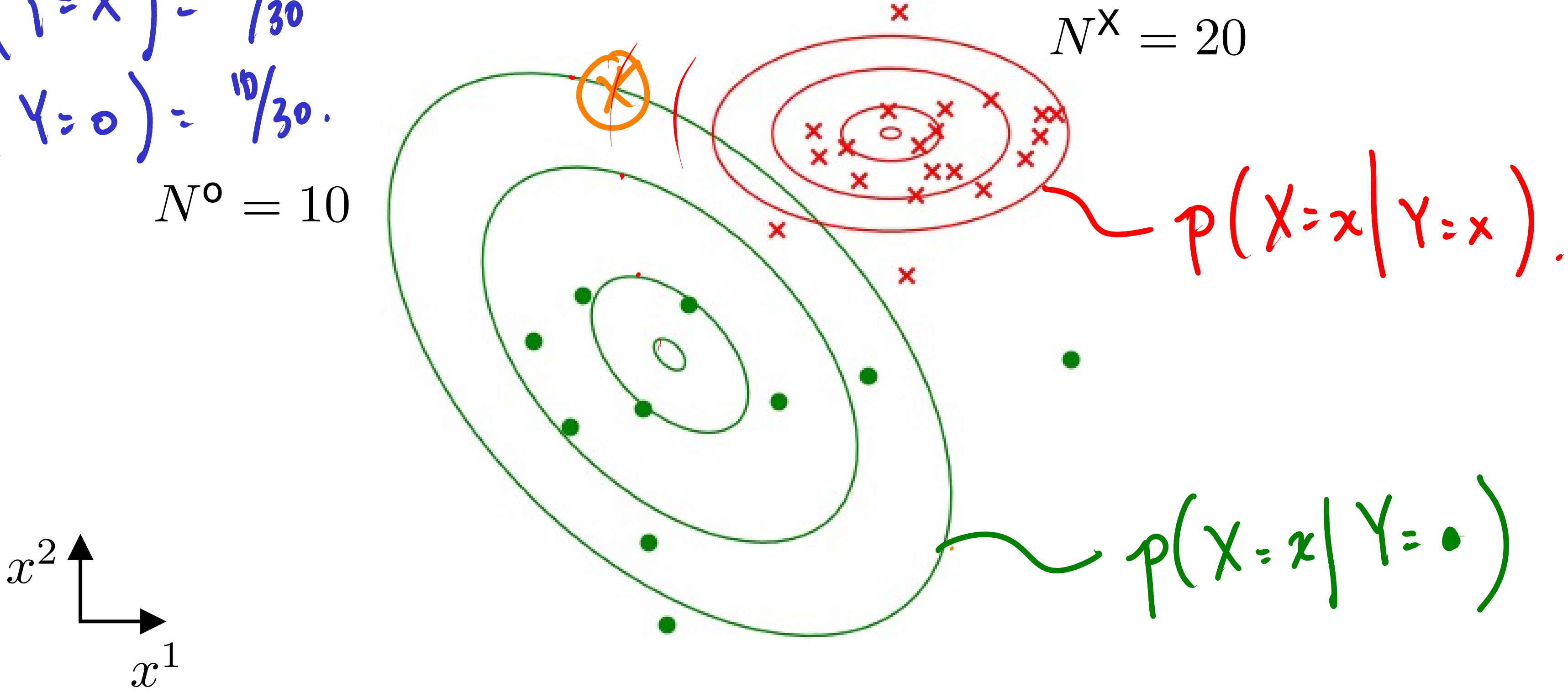
I Estimate a discrete distribution for every input $x \in \mathbb{R}^D$.

II Estimate a continuous distribution for every class + 1 discrete distr.

$$\hat{y} = \operatorname*{argmax}_{c \in \{c_1, \ldots, c_K\}} \underline{P(Y=c)}\,\underline{p(X^1=x^1, X^2=x^2 \mid Y=c)}$$

$P(Y=\times) = 20/30$

$P(Y=\circ) = 10/30.$

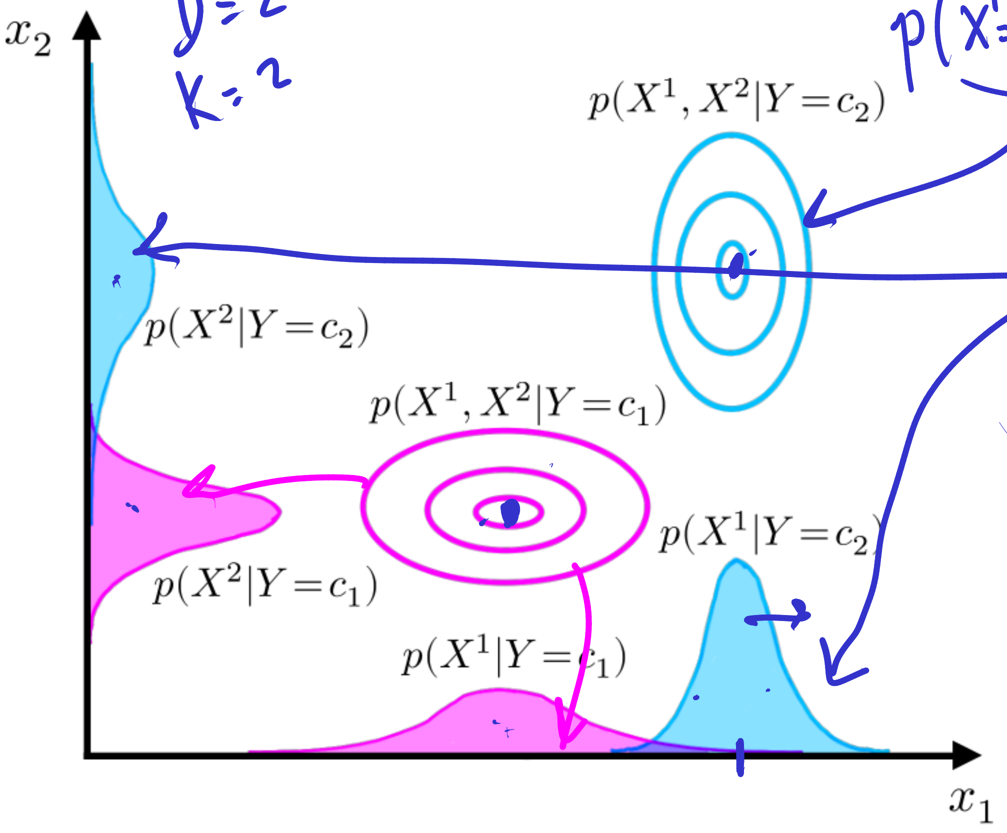


Naïve Bayes

Assumption: The components of the input are independent given the class.

$D = 2$
$K = 2$

$p(X^1 = x^1, X^2 = x^2 \mid Y = c) = p(X^1 = x^1 \mid Y = c)\, p(X^2 = x^2 \mid Y = c).$



... decomposed the multidimensional conditional distribution of the inputs into a product of univariate distributions

Original number of parameters.

K 2D Gaussians — D parameters for the mean
— $O(D^2)$ parameter for the covariance.

$\rightarrow K \cdot D^2$

Number of parameters after assumption.

$$K \cdot D \cdot 2.$$

$$O(KD^2) \longrightarrow O(KD).$$

Naïve Bayes

Assumption: The components of the input are independent given the class.

$$\hat{y} = \underset{c \in \{c_1, \ldots, c_K\}}{\text{argmax}} \quad P(Y=c) \prod_{d=1}^{D} p(X^d = x^d \mid Y = c)$$

$$= \underset{c \in \{c_1, \ldots, c_K\}}{\text{argmax}} \quad \log P(Y=c) + \sum_{d=1}^{D} \underbrace{\log p(X^d = x^d \mid Y = c)}_{\text{Gaussian}}$$

Gaussian Naïve Bayes

Assumption: The individual class-conditioned inputs are Gaussian

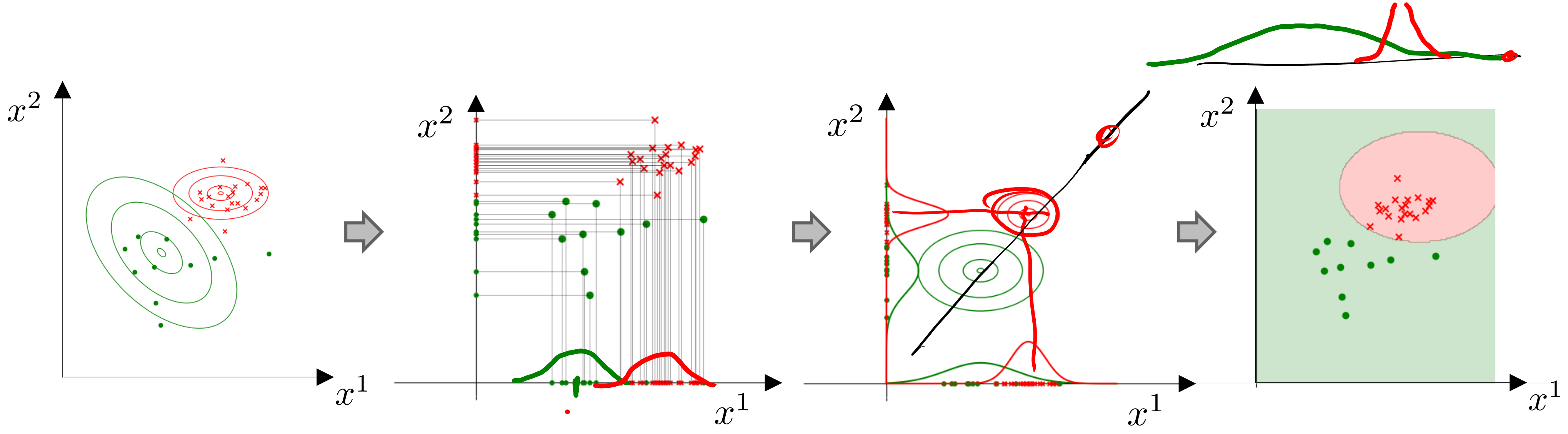
$$X^d = x^d | Y = c \ \sim \ \mathcal{N}(\mu_{d,c}, \sigma^2_{d,c}) \qquad \forall d, c$$

Training: Compute point estimates of the $\mu_{d,c}$'s and $\sigma^2_{d,c}$'s (and of the $P(Y=c)$'s)

Prediction: Choose the class that maximizes the posterior probability:

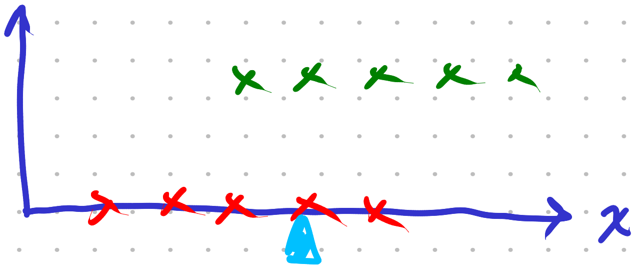
$$\hat{y} = \underset{c \in \{c_1, \ldots, c_K\}}{\text{argmax}} \quad \log P(Y=c) + \sum_{d=1}^{D} \log p(X^d = x^d \mid Y = c)$$

Example: Classifying two-class / 2D data with Gaussian NB



Recap:

$$x \in \mathbb{R}^D \longrightarrow \boxed{} \longrightarrow y \in \{c_1, \dots, c_K\}.$$



- Maximize accuracy.

$\Updownarrow$

- Minimize L_{01} ... problem: L_{01} cannot be minimized easily.

$\Updownarrow$

- $\underset{c}{\text{Maximize}}\ P(Y=c \mid X=x)$. ... Don't know $p(Y \mid X)$.

$\Updownarrow$ Bayes' rule.

$\underset{c}{\text{Maximize}}\ p(X=x \mid Y=c)\, P(Y=c)$. ... approximate these probabilities

↓ Naïve Bayes assumption. $p(X=x \mid Y=c) = \prod_{d=1}^{D} p(X^d = x^d \mid Y=c)$.

Naïve Bayes

Assume $p(X^d = x^d \mid Y=c)$ are Gaussian

Gaussian Naïve Bayes.

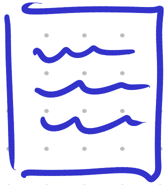
Inputs are label based.

⋮

text classification.

Text classification

Document



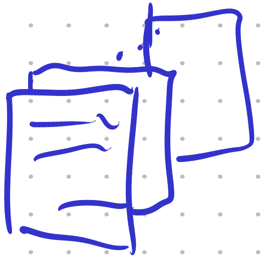
$$\xrightarrow{x} \boxed{h(x)} \longrightarrow \hat{y} \in \{c_1, \dots, c_k\}.$$

$\vdots$

{children's book, romance novel, scifi story ...}.

{positive, negative}.

Training data: "corpus"



How do we encode the document into D inputs. : Bag-of-words.

Naïve Bayes with labeled inputs

$x^1 \in \{a, b, c\}$

$x^2 \in \{\alpha, \beta, \gamma\}$

$y \in \{\bullet, \bullet, \bullet\}$

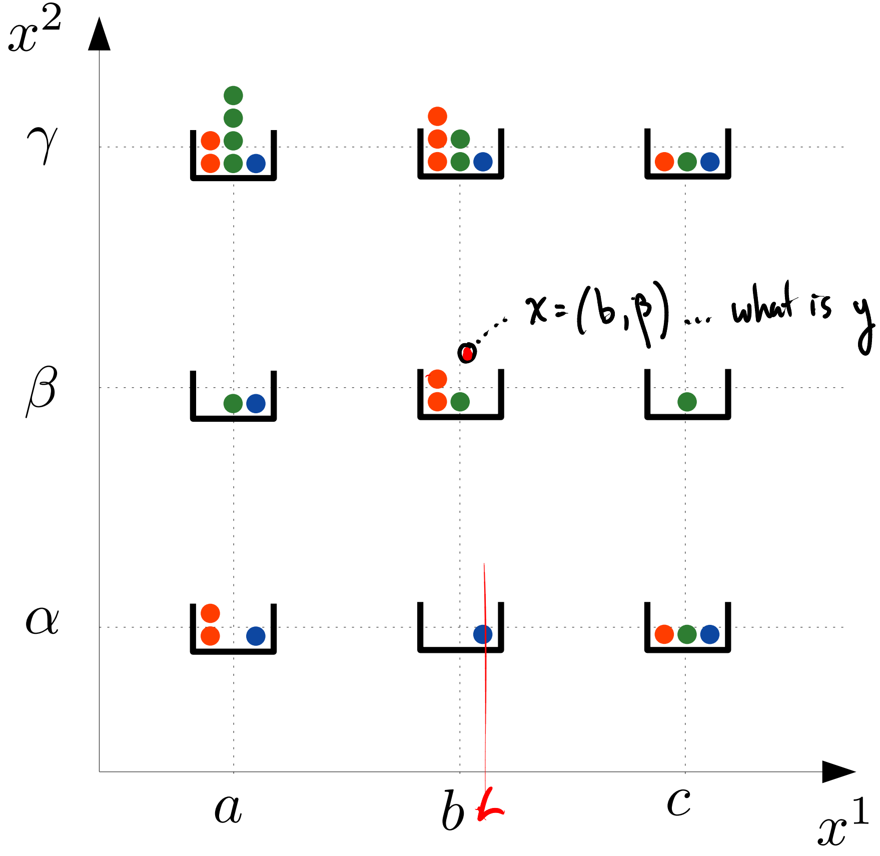
$P(Y = \bullet) \approx \dfrac{\#\text{ red dots}}{N}$

$= 11/29 = \hat{P}_R$

$P(Y = \bullet) = 11/29 = \hat{P}_G$

$P(Y = \bullet) = 7/29 = \hat{P}_B$

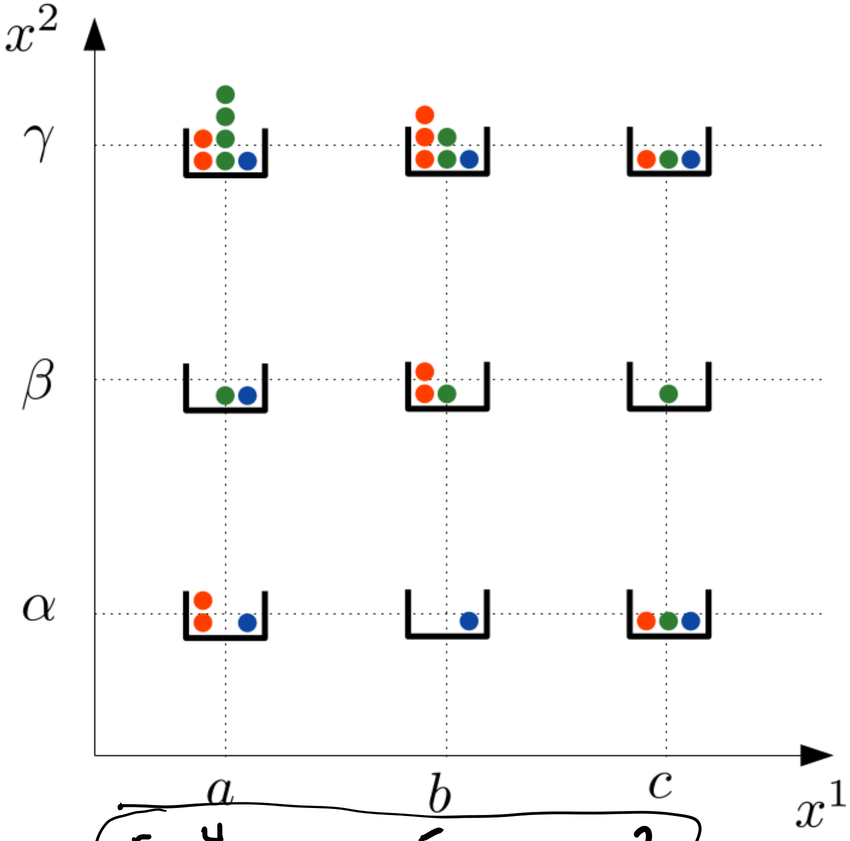
P



$x = (b, \beta)$ ... what is y? Compute $\hat{y}$

1

	r	g	b
γ	6	7	3
β	2	3	1
α	3	1	3



	a	b	c
r	4	5	2
g	5	3	3
b	3	2	2
	12	10	7

$$\hat{y} = \operatorname*{argmax}_c \underbrace{\log P(Y=c)}_{\checkmark} + \sum_{d=1}^{D} \log \underbrace{p(X^d = x^d \mid Y=c)} \quad \bigstar$$

$$p(X^1 = b \mid Y = \bullet) = 5/11$$

$$p(X^2 = \beta \mid Y = \bullet) = 2/11$$

$$\hat{y} = \operatorname{argmax}\Big(\underbrace{\log 11/29 + \log 5/11 + \log 2/11}_{\text{red}},\ \underbrace{\qquad\qquad\qquad}_{\text{green}},\ \underbrace{\qquad\qquad\qquad}_{\text{blue}} \Big)$$

Train the model: Use D_{Train} to comput

$$\hat{p}(Y=c)$$

$$\hat{p}(X^{d}=x^{d} \mid Y=c)$$

Evaluate/Prediction: evaluating argmax(log . . .) ☆

Naïve Bayes with labeled inputs

$$\hat{y} = \underset{c\in\{{\color{red}\bullet}{\color{green}\bullet}{\color{blue}\bullet}\}}{\text{argmax}} \; \log P(Y=c) + \sum_{d=1}^{D} \log p(X^d = x^d \mid Y=c)$$

$$= \underset{c\in\{{\color{red}\bullet}{\color{green}\bullet}{\color{blue}\bullet}\}}{\text{argmax}} \Big(\underbrace{\log 11/29 + \log 5/11 + \log 2/11}_{\text{red}},\; \underbrace{\log 11/29 + \log 3/11 + \log 3/11}_{\text{Green}},\; \underbrace{\log 7/29 + \log 2/7 + \log 1/7}_{\text{blue.}} \Big)$$

$$= \underset{c}{\text{argmax}} \left(\frac{\cancel{11}}{\cancel{29}}\,\frac{5}{\cancel{11}}\,\frac{2}{11},\; \frac{\cancel{11}}{\cancel{29}}\,\frac{3}{\cancel{11}}\,\frac{3}{11},\; \frac{\cancel{7}}{\cancel{29}}\,\frac{2}{\cancel{7}}\,\frac{1}{7} \right)$$

$$= \underset{c}{\text{argmax}} \left(\underset{\color{red}\bullet}{10/11},\; \underset{\color{green}\bullet}{9/11},\; \underset{\color{blue}\bullet}{2/7} \right) = {\color{red}\bullet}$$

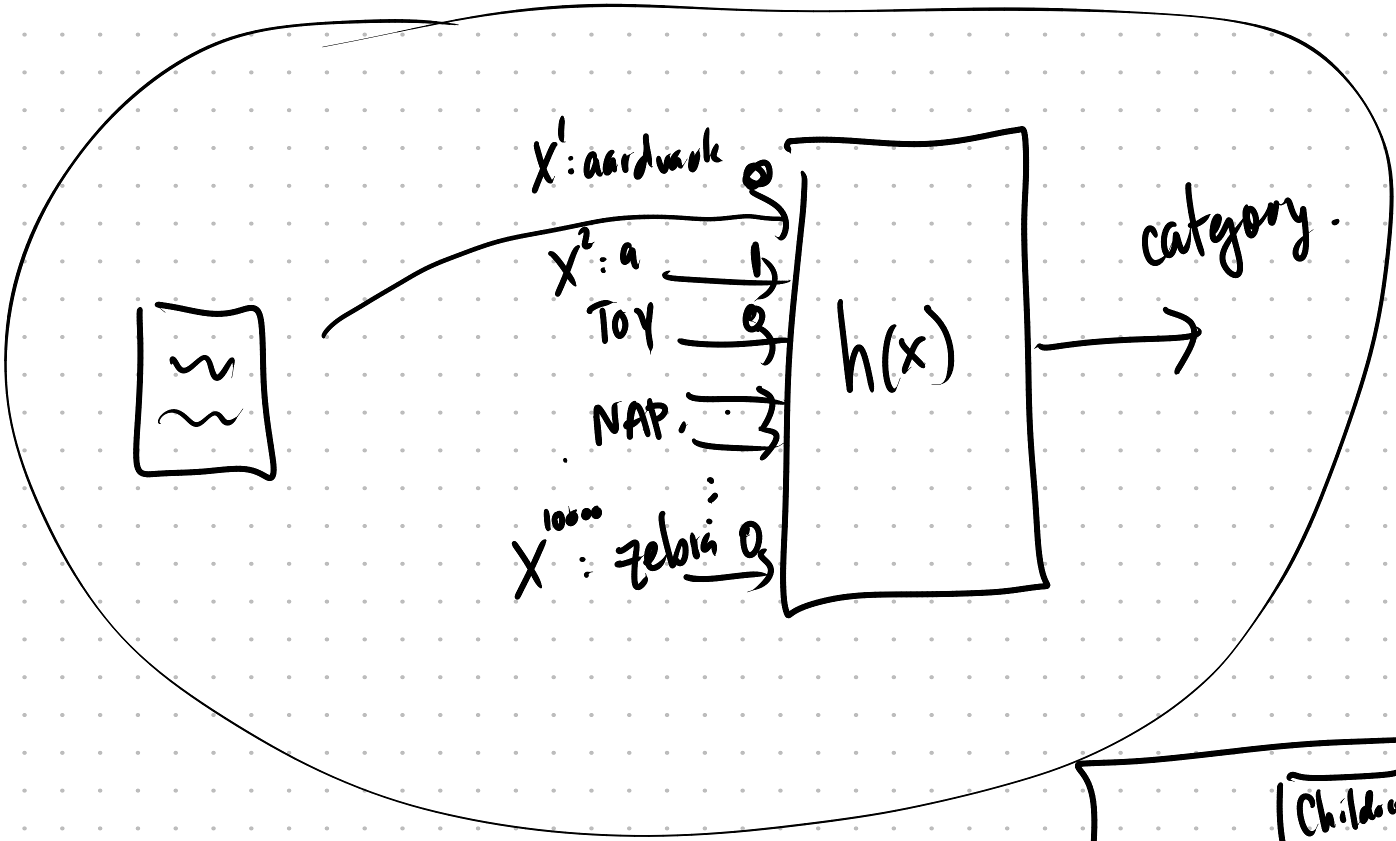
Example: Text classification

Problem

- A sample x is a *document* (article, blog entry, tweet, review, etc).
- Categories $\{c_1, c_2, \ldots, c_K\}$ (e.g. positive/negative, genre, etc).
- Training data (a *corpus*): $\mathcal{D} = \{(x_i, y_i)\}_N$

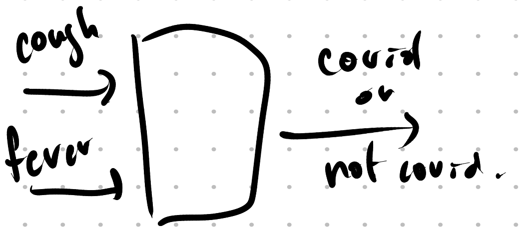
Input encoding: Vocabulary of D words.

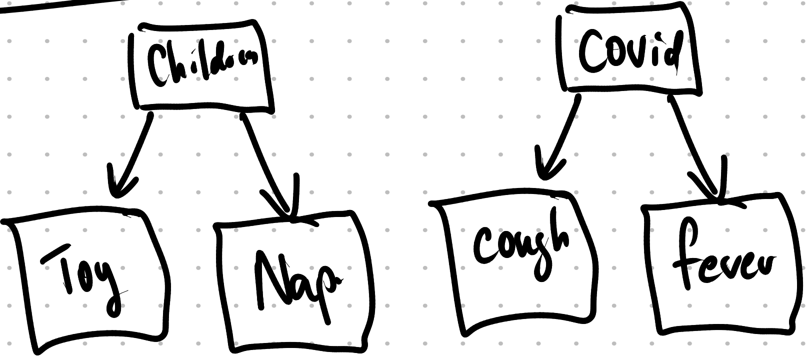
X^1 = a is present
X^2 = aardvark is present (all binary)
$\vdots$
$X^{10,000}$ = zebra is present



Aside on the Naïve Bayes assumption.

that the inputs are independent given the output:





Bag of words:

"y: target"

a	abacus	abandon	amazing	...	free	....	class
1	0	0	0	...	0	...	c1
1	0	0	0	...	0	...	c1
1	0	1	0	...	0	...	c3
1	0	0	0	...	0	...	c1
1	0	0	0	...	0	...	c2
1	0	0	0	...	0	...	c3
1	1	0	0	...	0	...	c1
1	0	0	0	...	0	...	c2
1	0	0	0	...	0	...	c1
1	0	0	1	...	0	...	c2

$$\hat{y} = \underset{c \in \{c_1, c_2, c_3\}}{\text{argmax}} \quad \underbrace{\log P(Y=c)}_{\checkmark} + \sum_{d=1}^{D} \underbrace{\log p(X^d = x^d \mid Y=c)}$$

vocabulary.

$$P(Y=c) \approx \frac{\text{\# docs of class } c}{\text{\# docs}} = \frac{N_c}{N} = \hat{p}_c$$

word d is present.

$$P(\overbrace{X^d = 1} \mid Y=c) \approx \frac{\text{\# docs of class } c \text{ with word } d}{\text{\# docs of class } c} = \boxed{\frac{N_{d,c}}{N_c} = \hat{p}_{d,c}}$$

word d is not present

$$P(\overbrace{X^d = 0} \mid Y=c) \approx 1 - \hat{p}_{d,c}$$

Problems:

1) Empty class: $N_c = 0 \quad \Rightarrow \quad N_{d,c} = 0 \quad \forall d \;\Rightarrow\; \hat{p}_{d,c}$ is undefined $\forall d$

2) New word: $N_{d,c}$ is undefined

Solution: Laplace smoothing

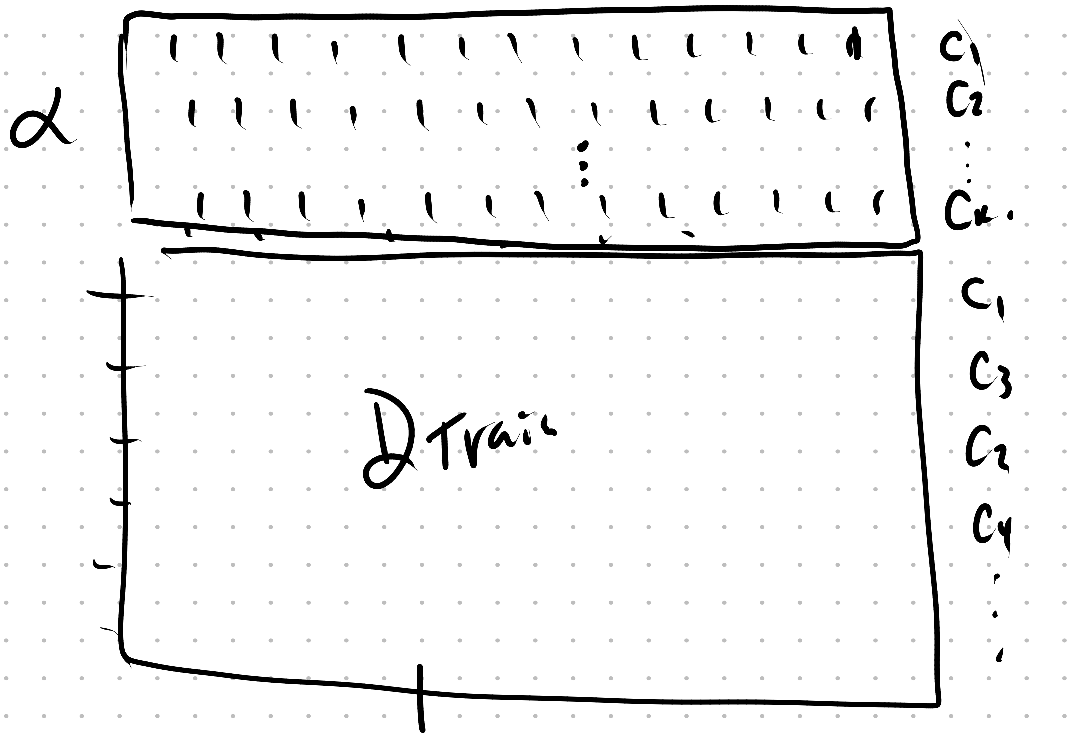
$$\boxed{\hat{p}_{d,c} = \frac{N_{d,c} + \alpha}{N_c + \alpha K}}$$

hyperparameter.

$\alpha > 0$

$$N_{d,c} \longrightarrow N_{d,c} + 1 \cdot \alpha$$

$$N_c \longrightarrow N_c + \underline{K} \cdot \alpha$$



Ndc are increased by 1.

All Nc's are increased by K.

Predict the class of this new document:

a	abacus	abandon	amazing	...	free	....
1	0	1	0	...	0	...

$$\hat{y} = \operatorname*{argmax}_c \; \log P(Y=c) + \sum_{d=1}^{D} \log P(X^d = x^d \mid Y=c)$$

$$\approx \operatorname*{argmax}_c \; \log \hat{p}_c + \sum_{d:x^d=1} \log \hat{p}_{d,c} + \sum_{d:x^d=0} \log(1-\hat{p}_{d,c})$$

word in the document

words not in the document.

Sorted table:

a	abacus	abandon	amazing	...	free	....	class
1	0	0	0	...	0	...	c1
1	0	0	0	...	0	...	c1
1	0	0	0	...	0	...	c1
1	1	0	0	...	0	...	c1
1	0	0	0	...	0	...	c1
1	0	0	0	...	0	...	c2
1	0	0	0	...	0	...	c2
1	0	0	1	...	0	...	c2
1	0	1	0	...	0	...	c3
1	0	0	0	...	0	...	c3