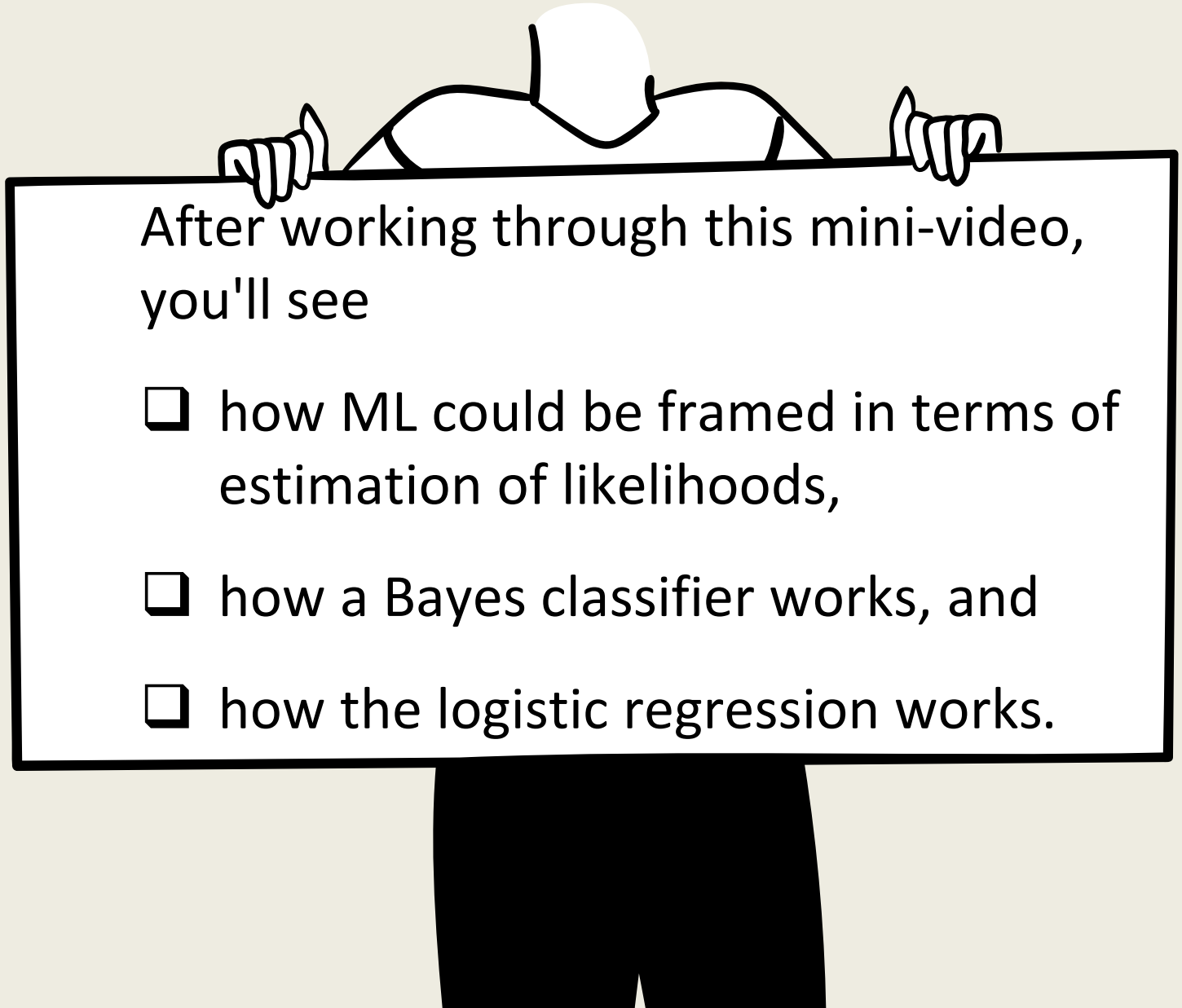
Probabilistic models

MACHINE LEARNING

Dr. Temitayo Olugbade

UNIVERSITY
OF SUSSEX

Learning outcome



After working through this mini-video, you'll see

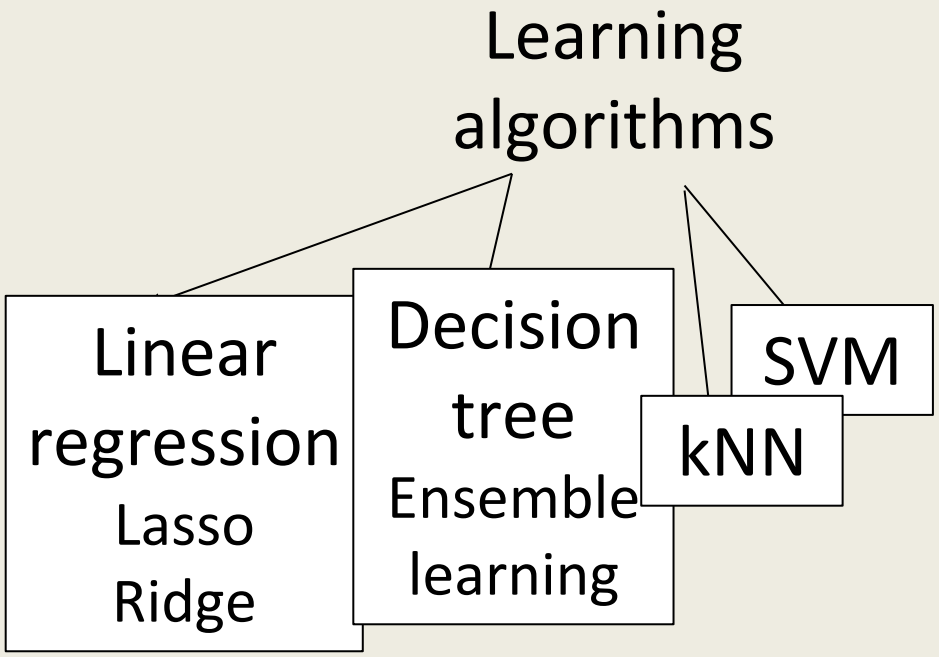
- ❑ how ML could be framed in terms of estimation of likelihoods,
- ❑ how a Bayes classifier works, and
- ❑ how the logistic regression works.

Mini-video outline

- ❑ Bayes classifier
- ❑ Logistic regression

Recall from previous mini-videos

- The goal of ML is to learn a **model** that takes in input $\boldsymbol{x}$ and gives label $\hat{\boldsymbol{y}}$ close to $\boldsymbol{y}$



- A model is characterised by its **learning algorithm**, **parameters**, **loss functions**, **hyperparameters**.

Classification inference formulations

- Formulation so far
  - Given input data $\left\{x_{i_1}, x_{i_2}, \cdots x_{i_D}\right\}, \forall d, d = 1, 2, \cdots D$
  - What is $\hat{y}_i$? $\hat{y}_i \subset \{c_k,\ \forall k, k = 1, 2, \cdots K\}$
- A different (but equivalent) formulation
  - Given data $\left\{x_{i_1}, x_{i_2}, \cdots x_{i_D}\right\}, \forall d, d = 1, 2, \cdots D$
  - What is $p\left(c_k | x_{i_1}, x_{i_2}, \cdots x_{i_D}\right), \forall k, k = 1, 2, \cdots K$?
  - $\hat{y}_i = c_k$, for $\underset{k}{\operatorname{argmax}}\, p\left(c_k | x_{i_1}, x_{i_2}, \cdots x_{i_D}\right)$

Bayes classifier

- **Bayes classifier**
- Logistic regression

Bayes' Rule

For two events *A* and *B*

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

where $P(B|A)$ = conditional probability of B given observation of A,
aka posterior probability of B

$P(B)$ = prior probability of B

$P(A|B)$ = the probability of A having occurred given observation of B
(i.e. the probability of generating A from the associated probability distribution with parameters conditioned on B)
aka likelihood of A

$P(A)$ = probability of A

Probabilistic model formulation

rewriting Bayes' rule for a set of possible classes $\{c_k, \quad \forall k, k = 1,2,\cdots,K\}$

$$P(Y = c_k|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|Y = c_k)P(Y = c_k)}{P(\boldsymbol{X})}$$

let's ignore $P(\boldsymbol{X})$ *since it's common* $\forall k$ *and so can be factored out*

$$P(Y = c_k|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|Y=c_k)P(Y=c_k)}{P(\boldsymbol{X})}$$

which is the same as

$$P(Y = c_k|\boldsymbol{X}) = \frac{P\left(X_1=x_{j_1},X_2=x_{j_2},\cdots,X_D=x_{j_D}|Y=c_k\right)P(Y=c_k)}{P(\boldsymbol{X})}$$

Bayes classifier

$$P(c_k|\boldsymbol{X}) = \frac{P\left(x_{j_1}, x_{j_2}, \cdots, x_{j_D}|c_k\right)P(c_k)}{P(\boldsymbol{X})} \forall k$$

where c_k = the kth class

K = total number of classes

$x_{j_1}, x_{j_2}, \cdots x_{j_D}$ = input data

Bayes classifier – Probability term 1

$$P(c_k|\boldsymbol{X}) = \frac{P\left(x_{j_1}, x_{j_2}, \cdots, x_{j_D}|c_k\right) P(c_k)}{P(\boldsymbol{X})} \forall k$$

$P(c_k)$ can be estimated from the training data as

$$P(c_k) = \frac{N_{c_k}}{N}$$

where N_{c_k} = number of instances of class c_k in the training data

N = total number of instances in the training data

Bayes classifier – Probability term 2

$$P(c_k|\boldsymbol{X}) = \frac{P\left(x_{j_1}, x_{j_2}, \cdots, x_{j_D}|c_k\right) P(c_k)}{P(\boldsymbol{X})} \forall k$$

$P\left(x_{j_1}, x_{j_2}, \cdots, x_{j_D}|c_k\right)$ can be estimated from the training data assuming a given distribution, e.g. Gaussian distribution

Bayes classifier – $P(\boldsymbol{X}|c_k)$

$$P\left(x_{j_1}, x_{j_2}, \cdots, x_{j_D}|c_k\right) =$$

$$= \frac{1}{(2\pi)^{D/2}} \frac{1}{\left|\boldsymbol{\Sigma}_{c_k}\right|^{1/2}} e^{-\frac{1}{2}\left(\boldsymbol{X}_{c_k}-\boldsymbol{\mu}_{c_k}\right)^T \boldsymbol{\Sigma}_{c_k}{}^{-1}\left(\boldsymbol{X}_{c_k}-\boldsymbol{\mu}_{c_k}\right)}$$

where

$$\boldsymbol{\mu}_{c_k} = \frac{1}{N_{c_k}} \sum_{n=1}^{N_{c_k}} \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_{c_k} = \frac{1}{N_{c_k}} \sum_{n=1}^{N_{c_k}} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right)^T \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right)$$

see math proof in last slide pages

Naïve Bayes classifier

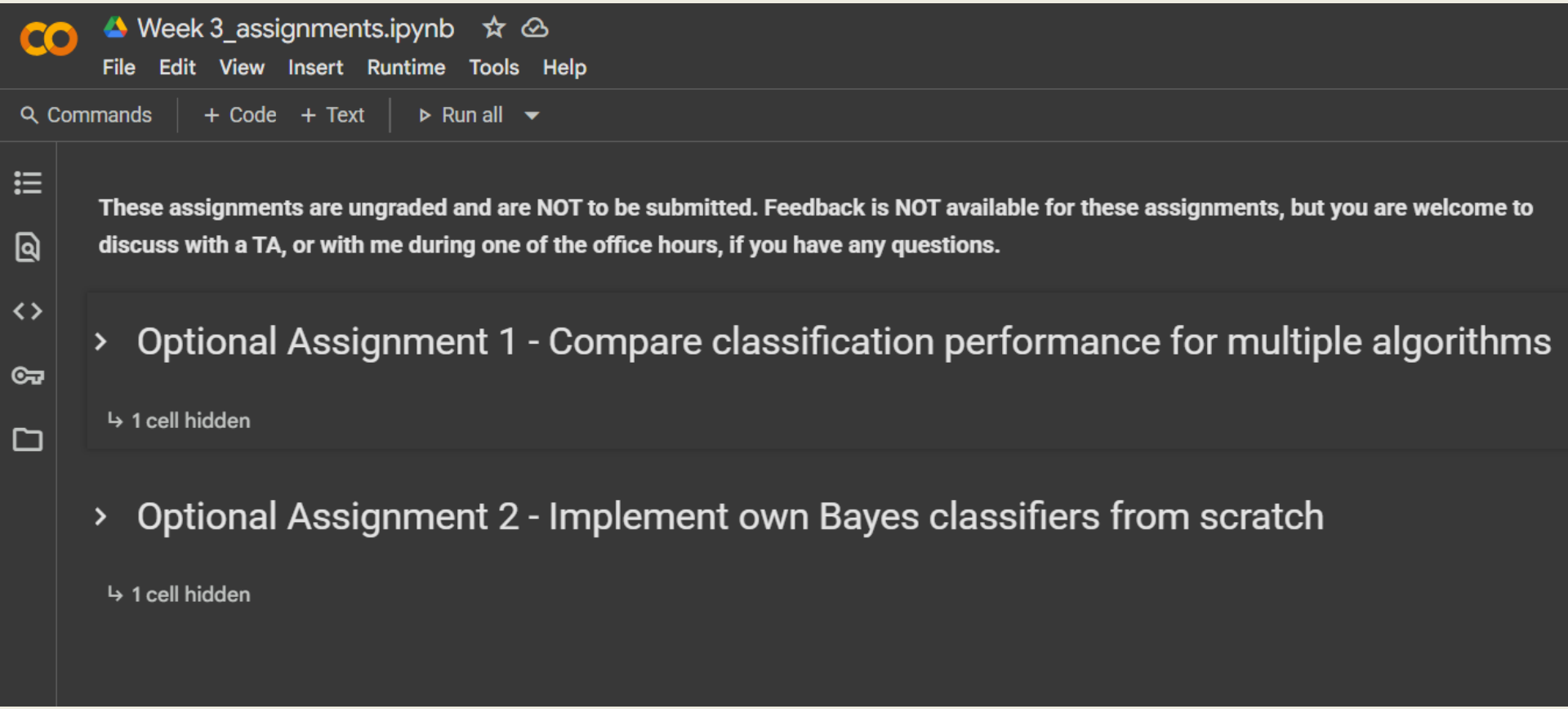
$$P(c_k|\boldsymbol{X}) = \frac{P\left(x_{j_1}, x_{j_2}, \cdots, x_{j_D}|c_k\right)P(c_k)}{P(\boldsymbol{X})} \forall k$$

Assuming that $x_{j_1}, x_{j_2}, \cdots x_{j_D}$ are conditionally independent given class c_k

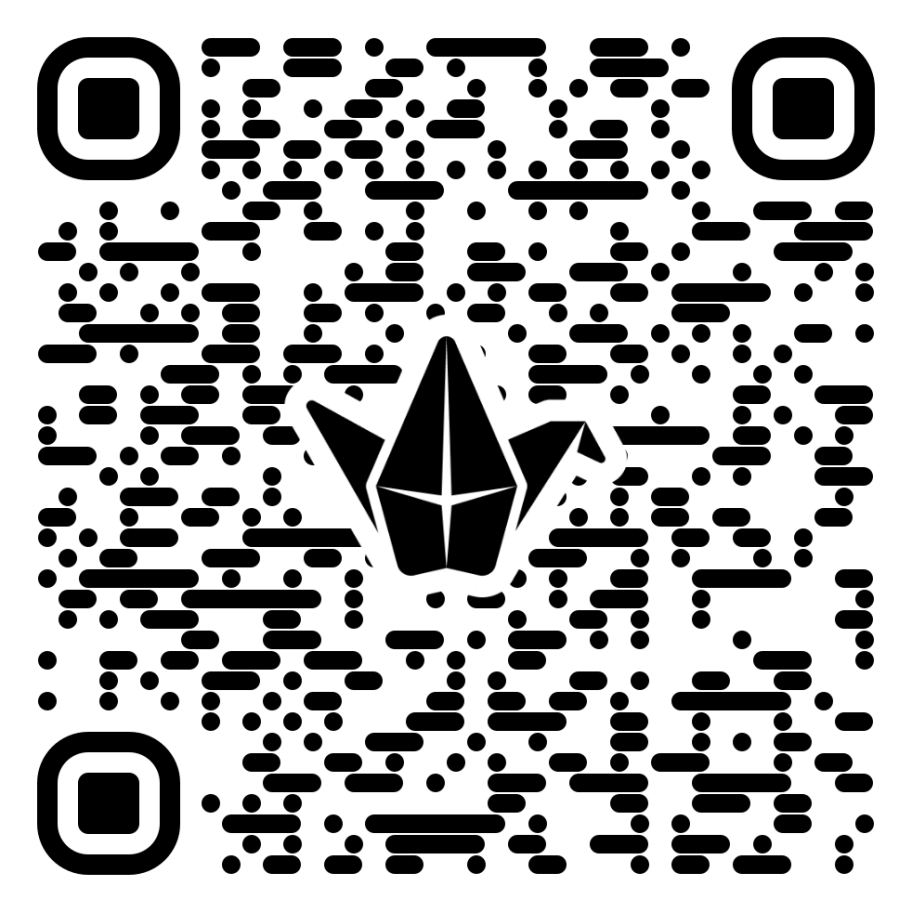
$$\Longrightarrow P(c_k|\boldsymbol{X}) = \frac{\prod_{d=1}^{D} P\left(x_{j_d}|c_k\right)P(c_k)}{P(\boldsymbol{X})}$$

- **Naïve**! – since this assumption is rarely true in the real world
- but the assumption simplifies $P\left(x_{j_1}, x_{j_2}, \cdots, x_{j_D}|c_k\right)$
- $P\left(x_{j_d}|c_k\right)$ can be estimated from the training data

Implement a Bayes classifier yourself



Any questions???



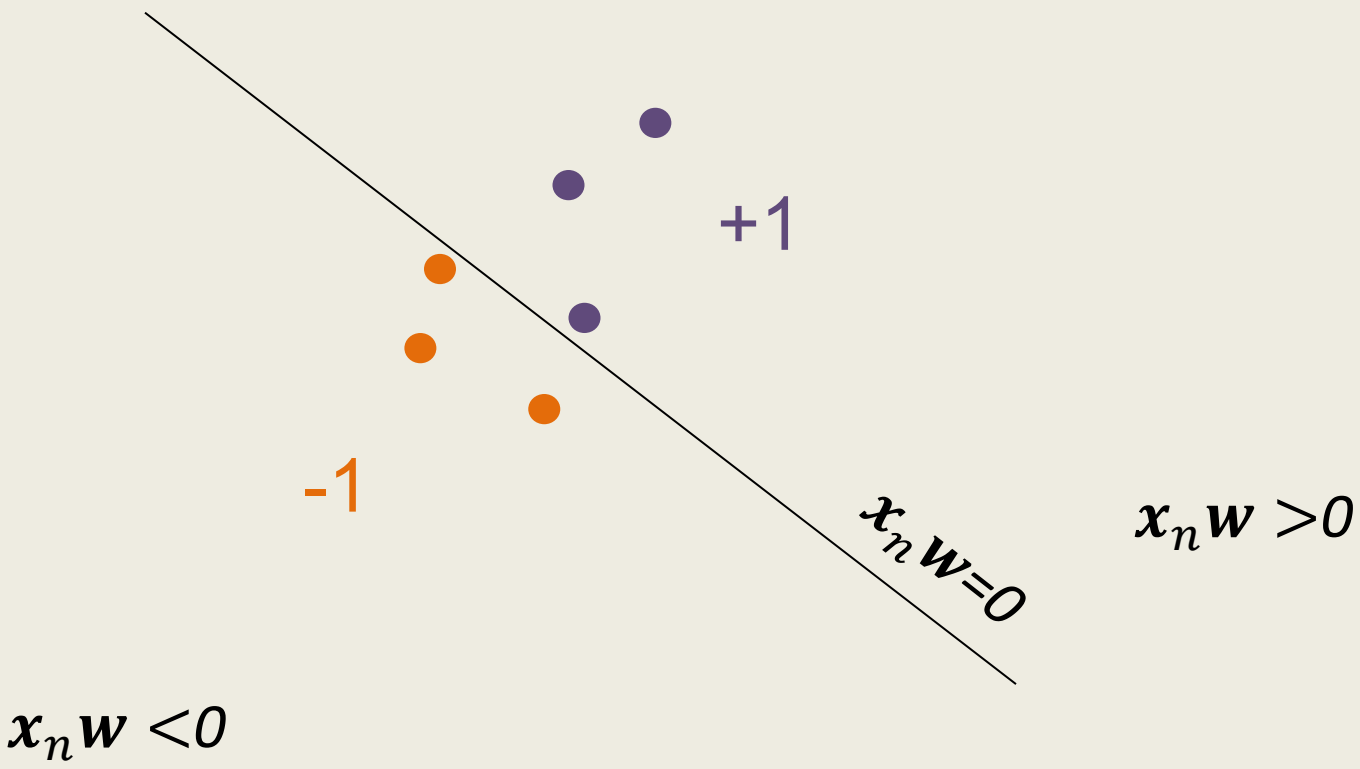
scan the QR code to ask questions

Logistic regression

- Bayes classifier

- **Logistic regression**

Re: Linear model (Classification)



The decision boundary is $x_n w = 0$

Re: Classification inference formulations

- Formulation so far
  - Given input data $\{x_{i_1}, x_{i_2}, \cdots x_{i_D}\}, \forall d, d = 1, 2, \cdots D$
  - What is $\hat{y}_i$? $\hat{y}_i \subset \{c_k, \ \forall k, k = 1, 2, \cdots K\}$
- A different (but equivalent) formulation
  - Given data $\{x_{i_1}, x_{i_2}, \cdots x_{i_D}\}, \forall d, d = 1, 2, \cdots D$
  - What is $p(c_k | x_{i_1}, x_{i_2}, \cdots x_{i_D}), \forall k, k = 1, 2, \cdots K$?
  - $\hat{y}_i = c_k$, for $\underset{k}{\operatorname{argmax}}\, p(c_k | x_{i_1}, x_{i_2}, \cdots x_{i_D})$

Decision boundaries (Binary classification)

- One formulation
  - Given data $\{x_{i_1}, x_{i_2}, \cdots x_{i_D}\}, \forall d, d = 1, 2, \cdots D$
  - What is $\hat{y}_i$? $\hat{y}_i \subset \{-1, 1\}$
  - $\boldsymbol{x_i w = 0}$ is the decision boundary
- A different (but equivalent) formulation
  - Given data $\{x_{i_1}, x_{i_2}, \cdots x_{i_D}\}, \forall d, d = 1, 2, \cdots D$
  - What is $p(c_k | x_{i_1}, x_{i_2}, \cdots x_{i_D}), \forall k, k = 1, 2$?
  - $\hat{y}_i = c_k$, for $\underset{k}{\operatorname{argmax}}\, p(c_k | x_{i_1}, x_{i_2}, \cdots x_{i_D})$
  - $\frac{p(Y=1|\boldsymbol{x}_i)}{p(Y=-1|\boldsymbol{x}_i)} = 1$ is the decision boundary

Logistic regression (LR)

$$\frac{P(Y = 1 \mid \boldsymbol{x})}{P(Y = 0 \mid \boldsymbol{x})} = 1$$

$$\Longrightarrow P(Y = 1 \mid \boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{xw}}}$$

the logistic regression (LR) model

see math proof on last slide pages

Decision boundaries (Binary classification)

- One formulation
  - Given data $\{x_{i_1}, x_{i_2}, \cdots x_{i_D}\}, \forall d, d = 1, 2, \cdots D$
  - What is $\hat{y}_i$? $\hat{y}_i \subset \{-1, 1\}$
  - $\boldsymbol{x_i w = 0}$ is the decision boundary
- A different (but equivalent) formulation
  - Given data $\{x_{i_1}, x_{i_2}, \cdots x_{i_D}\}, \forall d, d = 1, 2, \cdots D$
  - What is $p(c_k | x_{i_1}, x_{i_2}, \cdots x_{i_D}), \forall k, k = 1, 2$?
  - $\hat{y}_i = c_k$, for $\underset{k}{\operatorname{argmax}}\, p(c_k | x_{i_1}, x_{i_2}, \cdots x_{i_D})$
  - $\frac{p(Y=1|x_i)}{p(Y=-1|x_i)} = 1$ is the decision boundary

Re: Basic linear model for classification

$$f(\boldsymbol{x}) = \sigma(\boldsymbol{x}\boldsymbol{w} + b) = \hat{\boldsymbol{y}}$$

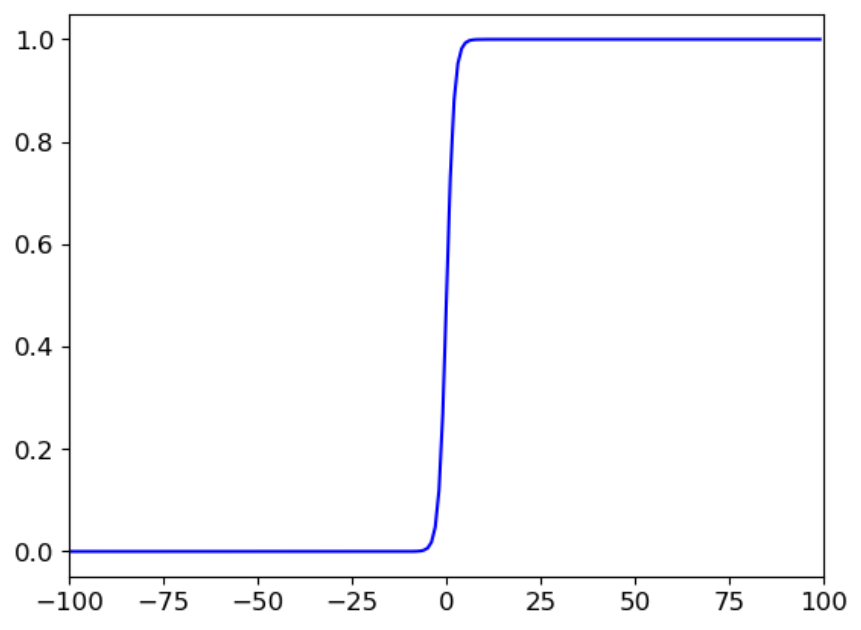
where

- $f(\cdot)$ – basic linear model
- $\boldsymbol{x}$ – features (or model input)
- $\hat{\boldsymbol{y}}$ – predicted labels/targets (or model output)
- $\boldsymbol{w}$, b – weights, bias (or model parameters)
- $\sigma(\cdot)$ – activation function (for discretizing real values)

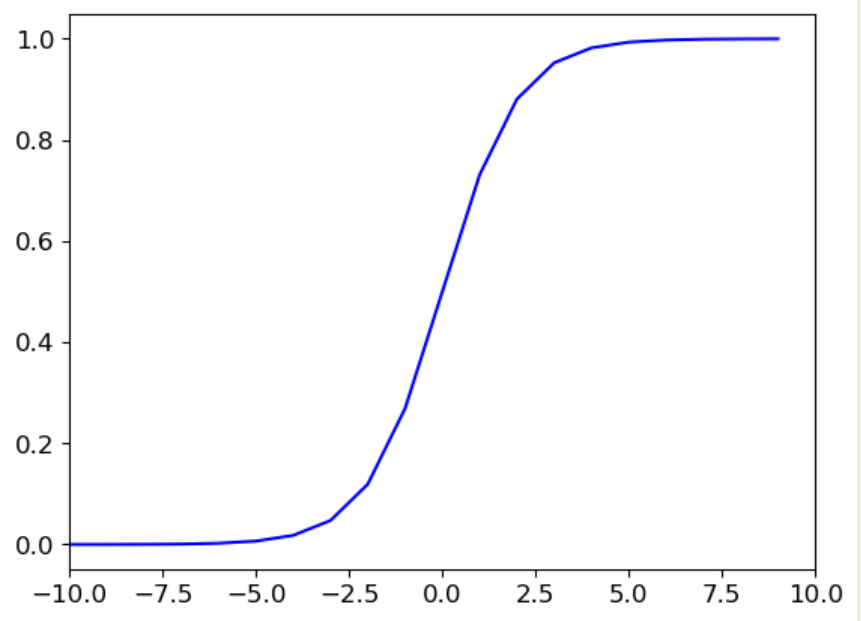
(Logistic) sigmoid activation function

$$f(\boldsymbol{x}) = \theta(\boldsymbol{xw} + b) = \frac{e^{\boldsymbol{xw}+b}}{e^{\boldsymbol{xw}+b} + 1} = \frac{1}{1 + e^{-(\boldsymbol{xw}+b)}}$$

$\theta(z)$



z



z

Optimal LR parameters (1)

- Likelihood based on a Bernoulli probability density function is **cross entropy loss** L_{CE}

$$L_{CE} = \sum_{n=1}^{N} -y_n \log f(\boldsymbol{x}_n) - (1 - y_n) \log\big(1 - f(\boldsymbol{x}_n)\big)$$

see math proof on last slide pages

- Model optimization involves maximizing likelihood (equivalent to finding the weights that minimize the loss)

$$\Rightarrow \sum_{n=1}^{N} \boldsymbol{x}_n (f(\boldsymbol{x}_n) - y_n) = 0$$

see math proof on last slide pages

Optimal LR parameters (2)

- There is no closed form solution for

$$\sum_{n=1}^{N} \boldsymbol{x}_n (f(\boldsymbol{x}_n) - y_n) = 0$$

  i.e. it cannot be solved analytically (to find the optimal $\boldsymbol{w}$)

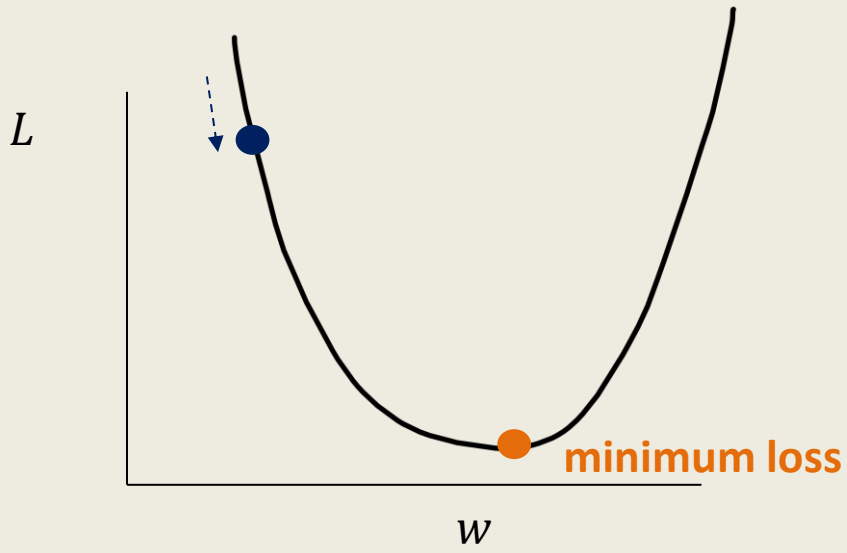
- The alternative is iterative solution using gradient descent algorithm

Gradient descent algorithm

- The optimal model parameters are at

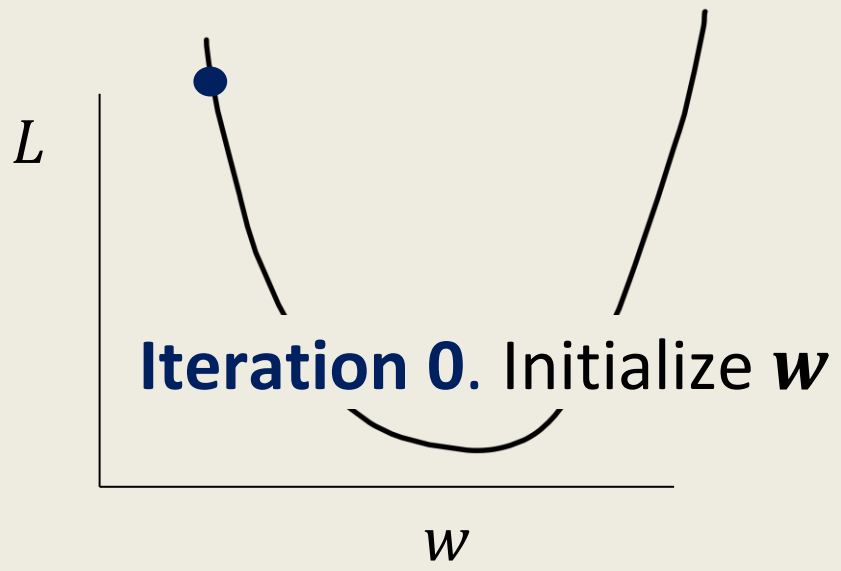
$$\frac{\partial L}{\partial \boldsymbol{w}} = 0$$

- i.e. in the direction of negative gradient (derivative) aka ‘gradient descent’

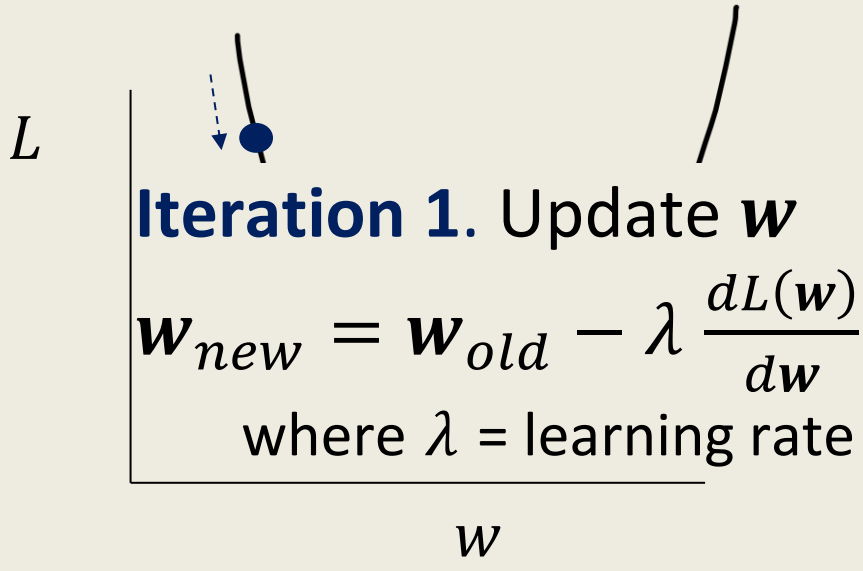


1D loss landscape for a simplified illustration

Descending iteratively



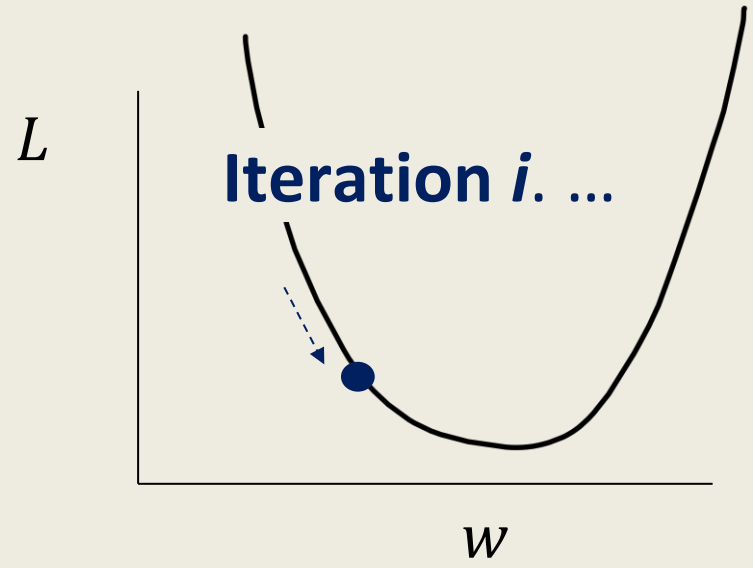
Iteration 0. Initialize $\boldsymbol{w}$



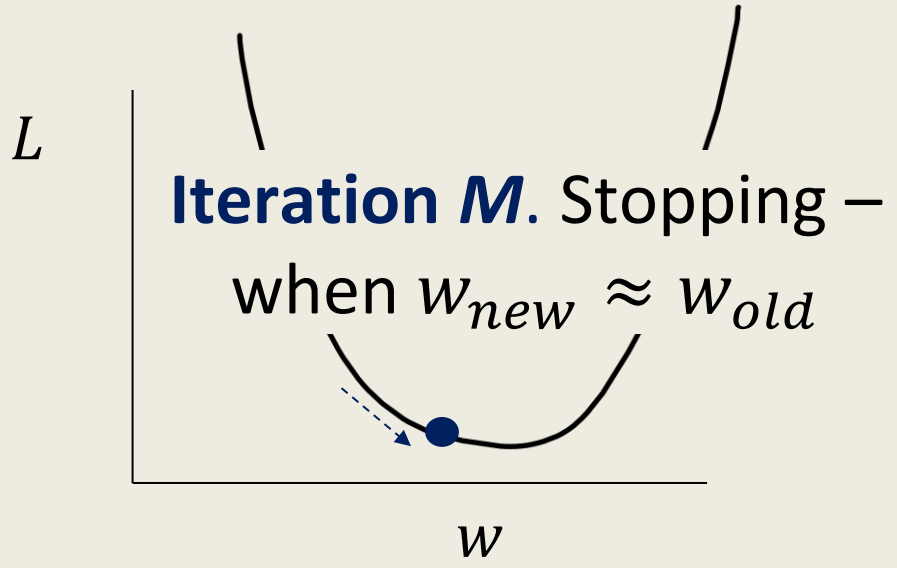
Iteration 1. Update $\boldsymbol{w}$

$$\boldsymbol{w}_{new} = \boldsymbol{w}_{old} - \lambda \frac{dL(\boldsymbol{w})}{d\boldsymbol{w}}$$

where λ = learning rate



Iteration *i*. …



Iteration *M*. Stopping – when $w_{new} \approx w_{old}$

Addressing LR overfitting

$$L_{LR} = \sum_{n=1}^{N} -y_n \log f(\boldsymbol{x}_n) - (1 - y_n) \log\big(1 - f(\boldsymbol{x})\big) + \alpha \frac{\|w\|^2}{2}$$

cross-entropy loss

regularization term

Summary

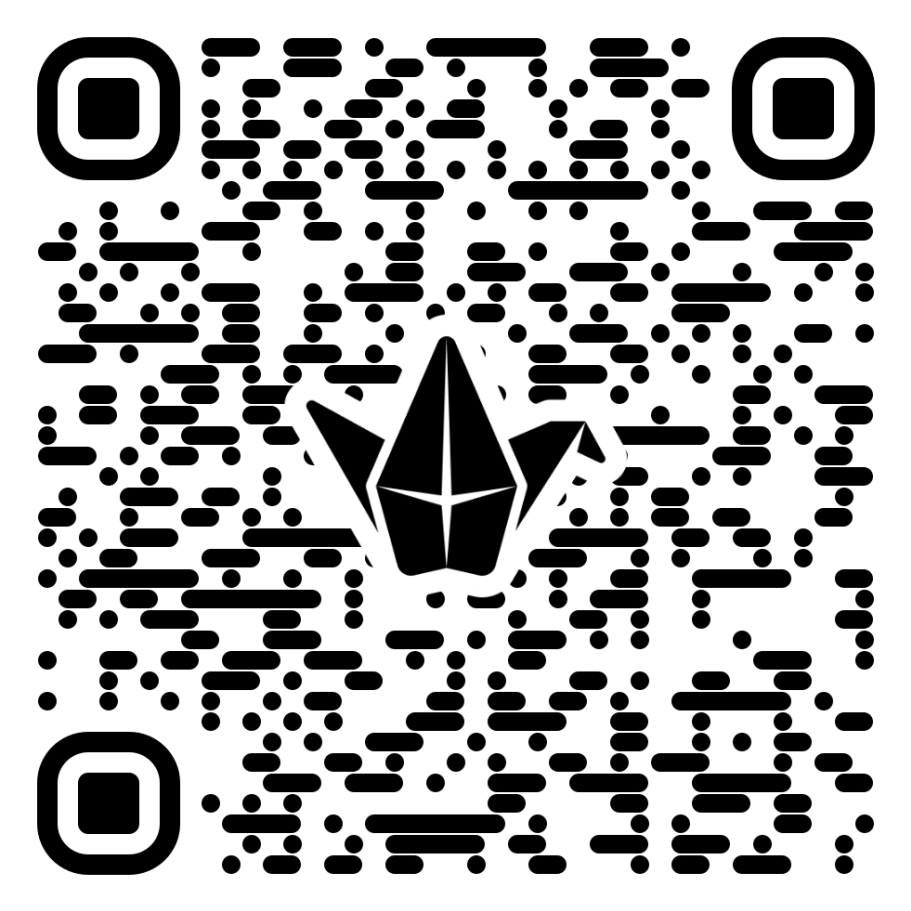
1. The **Bayes classifier** gives the probability of each class by updating the prior probability of the class with the likelihood of the observed data.

2. (Naïve) assumption of conditional independence in the Bayes model results in the **Naïve Bayes classifier**.

3. A **logistic regression** model is a classifier.

4. Its loss function is **cross-entropy loss**, and its parameters of classifier are optimized using **gradient descent algorithm**.

Any questions???



scan the QR code to ask questions

Math details and proofs

Bayes classifier – $P(\boldsymbol{X}|c_k)$ (MATH 0)

$$P(c_k|\boldsymbol{X}) = \frac{P\left(x_{j_1}, x_{j_2}, \cdots, x_{j_D}|c_k\right) P(c_k)}{P(\boldsymbol{X})} \forall k$$

$P\left(x_{j_1}, x_{j_2}, \cdots, x_{j_D}|c_k\right)$ can be estimated from the training data assuming a given distribution, e.g. Gaussian distribution $\mathcal{N}\left(\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right)$

$$\Longrightarrow P\left(x_{j_1}, x_{j_2}, \cdots, x_{j_D}|c_k\right) = \mathrm{E}\left[P\left(\boldsymbol{X}|\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right)\right]$$

where $\boldsymbol{\mu}_{c_k}$ = D-dimensional mean vector

$\boldsymbol{\Sigma}_{c_k}$ = $D \times D$ covariance matrix

$\mathrm{E}[\cdot]$ = expected value

Bayes classifier – $P(\boldsymbol{X}|c_k)$ (MATH 1)

- Assuming a Gaussian distribution $\mathcal{N}\left(\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right)$, the probability distribution parameters are $\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}$
- The optimal distribution parameters $\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}$ maximize the likelihood of the training data, equivalent to $\mathrm{E}\left[P\left(\boldsymbol{X}|\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right)\right]$
- Recall that optimum value of parameter θ for a function L is at

$$\frac{\partial L}{\partial \theta} = 0$$

Bayes classifier – $P(\boldsymbol{X}|c_k)$ (MATH 2)

The likelihood is Gaussian probability density function

$$P\left(\boldsymbol{X}|\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right) = \frac{1}{(2\pi)^{D/2}} \frac{1}{\left|\boldsymbol{\Sigma}_{c_k}\right|^{1/2}} e^{-\frac{1}{2}\left(\boldsymbol{X}_{c_k}-\boldsymbol{\mu}_{c_k}\right)^T \boldsymbol{\Sigma}_{c_k}{}^{-1}\left(\boldsymbol{X}_{c_k}-\boldsymbol{\mu}_{c_k}\right)}$$

taking log of both sides

$$\log P\left(\boldsymbol{X}|\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right)$$
$$= \log\left(\frac{1}{(2\pi)^{D/2}} \frac{1}{\left|\boldsymbol{\Sigma}_{c_k}\right|^{1/2}} e^{-\frac{1}{2}\left(\boldsymbol{X}_{c_k}-\boldsymbol{\mu}_{c_k}\right)^T \boldsymbol{\Sigma}_{c_k}{}^{-1}\left(\boldsymbol{X}_{c_k}-\boldsymbol{\mu}_{c_k}\right)}\right)$$

log of products is equivalent to sum of logs

$$\log P\left(\boldsymbol{X}|\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right)$$
$$= \log\left(\frac{1}{(2\pi)^{\frac{D}{2}}}\right) + \log\left(\frac{1}{\left|\boldsymbol{\Sigma}_{c_k}\right|^{\frac{1}{2}}}\right) + \log\left(e^{-\frac{\left(\boldsymbol{X}_{c_k}-\boldsymbol{\mu}_{c_k}\right)^T \boldsymbol{\Sigma}_{c_k}{}^{-1}\left(\boldsymbol{X}_{c_k}-\boldsymbol{\mu}_{c_k}\right)}{2}}\right)$$

Bayes classifier – $P(\boldsymbol{X}|c_k)$ (MATH 3)

$$\log P\left(\boldsymbol{X}|\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right)$$

$$= \log\left(\frac{1}{(2\pi)^{\frac{D}{2}}}\right) + \log\left(\frac{1}{\left|\boldsymbol{\Sigma}_{c_k}\right|^{\frac{1}{2}}}\right) + \log\left(e^{-\frac{\left(\boldsymbol{X}_{c_k}-\boldsymbol{\mu}_{c_k}\right)^T \Sigma_{c_k}{}^{-1}\left(\boldsymbol{X}_{c_k}-\boldsymbol{\mu}_{c_k}\right)}{2}}\right)$$

applying log rules

$$\log P\left(\boldsymbol{X}|\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right)$$

$$= -N_{c_k}\frac{D}{2}\log(2\pi) - N_{c_k}\frac{1}{2}\log\left|\boldsymbol{\Sigma}_{c_k}\right|$$

$$-\frac{1}{2}\sum_{n=1}^{N_{c_k}}\left(x_n - \boldsymbol{\mu}_{c_k}\right)^T \boldsymbol{\Sigma}_{c_k}{}^{-1}\left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right)$$

Bayes classifier – $P(\boldsymbol{X}|c_k)$ (MATH 4)

$$\Rightarrow \frac{\partial}{\partial \boldsymbol{\mu}} \log P\left(\boldsymbol{X}|\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right) = 0$$

substituting for $P\left(\boldsymbol{X}|\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right)$

$$\Rightarrow \frac{\partial}{\partial \boldsymbol{\mu}} \left(\frac{-N_{c_k} D}{2} \log(2\pi) - \frac{N_{c_k}}{2} \log|\boldsymbol{\Sigma}_{c_k}| \right.$$

$$\left. - \frac{1}{2} \sum_{n=1}^{N_{c_k}} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right)^T \boldsymbol{\Sigma}_{c_k}{}^{-1} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right) \right) = 0$$

removing terms without $\boldsymbol{\mu}$ based on differentiation rules

$$\Rightarrow \frac{\partial}{\partial \boldsymbol{\mu}} \left(- \frac{1}{2} \sum_{n=1}^{N_{c_k}} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right)^T \boldsymbol{\Sigma}_{c_k}{}^{-1} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right) \right) = 0$$

Bayes classifier – $P(\boldsymbol{X}|c_k)$ (MATH 5)

$$\frac{\partial}{\partial \boldsymbol{\mu}}\left(-\frac{1}{2}\sum_{n=1}^{N_{c_k}}\left(\boldsymbol{x}_n-\boldsymbol{\mu}_{c_k}\right)^T\boldsymbol{\Sigma}_{c_k}{}^{-1}\left(\boldsymbol{x}_n-\boldsymbol{\mu}_{c_k}\right)\right)=0$$

expanding the product

$$\frac{\partial}{\partial \boldsymbol{\mu}}\left(-\frac{1}{2}\sum_{n=1}^{N_{c_k}}\boldsymbol{x}_n{}^T\boldsymbol{\Sigma}_{c_k}{}^{-1}\boldsymbol{x}_n-2\boldsymbol{\mu}_{c_k}{}^T\boldsymbol{\Sigma}_{c_k}{}^{-1}\boldsymbol{x}_n+\boldsymbol{\mu}_{c_k}{}^T\boldsymbol{\Sigma}_{c_k}{}^{-1}\boldsymbol{\mu}_{c_k}\right)=0$$

further removing terms without $\boldsymbol{\mu}$ based on differentiation rules

$$\frac{\partial}{\partial \boldsymbol{\mu}}\left(-\frac{1}{2}\sum_{n=1}^{N_{c_k}}-2\boldsymbol{\mu}_{c_k}{}^T\boldsymbol{\Sigma}_{c_k}{}^{-1}\boldsymbol{x}_n+\boldsymbol{\mu}_{c_k}{}^T\boldsymbol{\Sigma}_{c_k}{}^{-1}\boldsymbol{\mu}_{c_k}\right)=0$$

Bayes classifier – $P(\boldsymbol{X}|c_k)$ (MATH 6)

$$\frac{\partial}{\partial \boldsymbol{\mu}}\left(-\frac{1}{2}\sum_{n=1}^{N_{c_k}} -2{\boldsymbol{\mu}_{c_k}}^T {\boldsymbol{\Sigma}_{c_k}}^{-1} \boldsymbol{x}_n + {\boldsymbol{\mu}_{c_k}}^T {\boldsymbol{\Sigma}_{c_k}}^{-1} \boldsymbol{\mu}_{c_k}\right) = 0$$

applying differentiation

$$\sum_{n=1}^{N_{c_k}} {\boldsymbol{\Sigma}_{c_k}}^{-1} \boldsymbol{x}_n - {\boldsymbol{\Sigma}_{c_k}}^{-1} \boldsymbol{\mu}_{c_k} = 0$$

collecting terms with $\boldsymbol{\mu}_{c_k}$ to the right hand side

$$\sum_{n=1}^{N_{c_k}} {\boldsymbol{\Sigma}_{c_k}}^{-1} \boldsymbol{x}_n = N_{c_k} {\boldsymbol{\Sigma}_{c_k}}^{-1} \boldsymbol{\mu}_{c_k}$$

Bayes classifier – $P(\boldsymbol{X}|c_k)$ (MATH 7)

$$\sum_{n=1}^{N_{c_k}} \boldsymbol{\Sigma}_{c_k}{}^{-1}\, \boldsymbol{x}_n = N_{c_k} \boldsymbol{\Sigma}_{c_k}{}^{-1} \boldsymbol{\mu}_{c_k}$$

dividing through by $\Sigma_{c_k}{}^{-1}$

$$\sum_{n=1}^{N_{c_k}} \boldsymbol{x}_n = N_{c_k} \boldsymbol{\mu}_{c_k}$$

making μ_{c_k} *the subject of the equation*

$$\Longrightarrow \boldsymbol{\mu}_{c_k} = \frac{1}{N_{c_k}} \sum_{n=1}^{N_{c_k}} \boldsymbol{x}_n$$

Bayes classifier – $P(\boldsymbol{X}|c_k)$ (MATH 8)

$$\Longrightarrow \frac{\partial}{\partial \boldsymbol{\Sigma}} \log P\left(\boldsymbol{X}|\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right) = 0$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} \left(\frac{-N_{c_k} D}{2} \log(2\pi) - \frac{N_{c_k}}{2} \log|\boldsymbol{\Sigma}_{c_k}| - \frac{1}{2} \sum_{n=1}^{N_{c_k}} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right)^T \boldsymbol{\Sigma}_{c_k}{}^{-1} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right) \right) = 0$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} \left(-\frac{N_{c_k}}{2} \log|\boldsymbol{\Sigma}_{c_k}| - \frac{1}{2} \sum_{n=1}^{N_{c_k}} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right)^T \boldsymbol{\Sigma}_{c_k}{}^{-1} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right) \right) = 0$$

Bayes classifier – $P(\boldsymbol{X}|c_k)$ (MATH 9)

$$\frac{\partial}{\partial \boldsymbol{\Sigma}}\left(-\frac{N_{c_k}}{2}\log|\boldsymbol{\Sigma}_{c_k}|\right)+\frac{\partial}{\partial \boldsymbol{\Sigma}}\left(\frac{1}{2}\sum_{n=1}^{N_{c_k}}\left(\boldsymbol{x}_n-\boldsymbol{\mu}_{c_k}\right)^T\boldsymbol{\Sigma}_{c_k}{}^{-1}\left(\boldsymbol{x}_n-\boldsymbol{\mu}_{c_k}\right)\right)$$
$$=0$$

$$-\frac{N_{c_k}}{2}\boldsymbol{\Sigma}_{c_k}{}^{-1}+\frac{1}{2}\sum_{n=1}^{N_{c_k}}\left(\boldsymbol{x}_n-\boldsymbol{\mu}_{c_k}\right)^T\boldsymbol{\Sigma}_{c_k}{}^{-2}\left(\boldsymbol{x}_n-\boldsymbol{\mu}_{c_k}\right)=0$$

$$-\frac{N_{c_k}}{2}\boldsymbol{\Sigma}_{c_k}{}^{-1}=-\frac{1}{2}\sum_{n=1}^{N_{c_k}}\left(\boldsymbol{x}_n-\boldsymbol{\mu}_{c_k}\right)^T\boldsymbol{\Sigma}_{c_k}{}^{-2}\left(\boldsymbol{x}_n-\boldsymbol{\mu}_{c_k}\right)$$

$$\Rightarrow \boldsymbol{\Sigma}_{c_k}=\frac{1}{N_{c_k}}\sum_{n=1}^{N_{c_k}}\left(\boldsymbol{x}_n-\boldsymbol{\mu}_{c_k}\right)^T\left(\boldsymbol{x}_n-\boldsymbol{\mu}_{c_k}\right)$$

Bayes classifier – $P(\boldsymbol{X}|c_k)$ (MATH 10)

$$P\left(x_{j_1}, x_{j_2}, \cdots, x_{j_D}|c_k\right) = P\left(\boldsymbol{X}|\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right)$$

where

$$P\left(\boldsymbol{X}|\boldsymbol{\mu}_{c_k}, \boldsymbol{\Sigma}_{c_k}\right) = \frac{1}{(2\pi)^{D/2}} \frac{1}{\left|\boldsymbol{\Sigma}_{c_k}\right|^{1/2}} e^{-\frac{1}{2}\left(\boldsymbol{X}_{c_k}-\boldsymbol{\mu}_{c_k}\right)^T \boldsymbol{\Sigma}_{c_k}{}^{-1}\left(\boldsymbol{X}_{c_k}-\boldsymbol{\mu}_{c_k}\right)}$$

$$\boldsymbol{\mu}_{c_k} = \frac{1}{N_{c_k}} \sum_{n=1}^{N_{c_k}} \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_{c_k} = \frac{1}{N_{c_k}} \sum_{n=1}^{N_{c_k}} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right)^T \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{c_k}\right)$$

Logistic regression (LR) – MATH PROOF (1)

$$\frac{P(Y=1 \mid \boldsymbol{x})}{P(Y=0 \mid \boldsymbol{x})} = 1$$

taking log of both sides

$$\log\left(\frac{P(Y=1 \mid \boldsymbol{x})}{P(Y=0 \mid \boldsymbol{x})}\right) = 0$$

substituting for $\boldsymbol{xw} = 0$ on the right hand side

$$\log\left(\frac{P(Y=1 \mid \boldsymbol{x})}{P(Y=0 \mid \boldsymbol{x})}\right) = \boldsymbol{xw}$$

applying probability rule

$$\log\left(\frac{P(Y=1 \mid \boldsymbol{x})}{1 - P(Y=1 \mid \boldsymbol{x})}\right) = \boldsymbol{xw}$$

Logistic regression (LR) – MATH PROOF (2)

$$\log\left(\frac{P(Y=1 \mid \boldsymbol{x})}{1-P(Y=1 \mid \boldsymbol{x})}\right) = \boldsymbol{xw}$$

taking exponent function of both sides

$$\frac{P(Y=1 \mid \boldsymbol{x})}{1-P(Y=1 \mid \boldsymbol{x})} = e^{\boldsymbol{xw}}$$

cross-multiplying

$$P(Y=1 \mid \boldsymbol{x}) = e^{\boldsymbol{xw}} - e^{\boldsymbol{xw}}\left(P(Y=1 \mid \boldsymbol{x})\right)$$

making $P(Y=1 \mid \boldsymbol{x})$ the subject of the formula

$$P(Y=1 \mid \boldsymbol{x}) = \frac{e^{\boldsymbol{xw}}}{1+e^{\boldsymbol{xw}}}$$

Logistic regression (LR) – MATH PROOF (3)

$$P(Y = 1 \mid \boldsymbol{x}) = \frac{e^{\boldsymbol{xw}}}{1 + e^{\boldsymbol{xw}}}$$

dividing numerator and denominator by $e^{\boldsymbol{xw}}$

$$P(Y = 1 \mid \boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{xw}}}$$

the logistic regression (LR) model

Optimal LR parameters (1) – PROOF (a)

- Consider binary classification and so, assuming Bernoulli distribution $Bern(\theta)$

- Likelihood based on a Bernoulli probability density function is

$$P(Y|\boldsymbol{X},\boldsymbol{w}) = \prod_{n=1}^{N} P(Y|\boldsymbol{x}_n)^{y_n}\left(1 - P(Y|\boldsymbol{x}_n)\right)^{1-y_n}$$

$$= \prod_{n=1}^{N} f(\boldsymbol{x}_n)^{y_n}\left(1 - f(\boldsymbol{x}_n)\right)^{1-y_n}$$

- Optimization involves maximizing likelihood $P(Y|\boldsymbol{X},\boldsymbol{w})$ equivalent to minimizing negative log likelihood

Optimal LR parameters (1) – PROOF (b)

$$\frac{\partial}{\partial \boldsymbol{w}}\left(-\log\left(P(Y|\boldsymbol{X},\boldsymbol{w})\right)\right) = 0$$

substituting for the formula of the likelihood

$$\frac{\partial}{\partial \boldsymbol{w}}\left(-\log\left(\prod_{n=1}^{N} f(\boldsymbol{x}_n)^{y_n}\left(1-f(\boldsymbol{x}_n)\right)^{1-y_n}\right)\right) = 0$$

applying logarithm rules

$$\frac{\partial}{\partial \boldsymbol{w}}\left(\sum_{n=1}^{N} -y_n \log f(\boldsymbol{x}_n) - (1-y_n)\log\left(1-f(\boldsymbol{x}_n)\right)\right) = 0$$

Optimal LR parameters (1) – PROOF (c)

$$\frac{\partial}{\partial \boldsymbol{w}}\left(\sum_{n=1}^{N} -y_n \log f(\boldsymbol{x}_n) - (1-y_n)\log\big(1-f(\boldsymbol{x}_n)\big)\right)$$

where $f(\boldsymbol{x}_n, \boldsymbol{w}) = \frac{1}{1+e^{-\boldsymbol{\theta}}}$ and $\boldsymbol{\theta} = \boldsymbol{x}_n \boldsymbol{w}$

$$\frac{\partial L}{\partial \boldsymbol{w}} = \frac{\partial L}{\partial f(\boldsymbol{\theta})} \cdot \frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot \frac{\partial \theta}{\partial \boldsymbol{w}}$$

where $L = \sum_{n=1}^{N} -y_n \log f(\boldsymbol{x}_n) - (1-y_n)\log\big(1-f(\boldsymbol{x}_n)\big)$

cross entropy loss

Optimal LR parameters (1) – PROOF (d)

$$\frac{\partial}{\partial \boldsymbol{w}}\left(\sum_{n=1}^{N} -y_n \log f(\boldsymbol{x}_n) - (1-y_n)\log\big(1-f(\boldsymbol{x}_n)\big)\right) = 0$$

applying rule of derivative of a function of a function

$$\sum_{n=1}^{N} -\left(\frac{y_n}{f(\boldsymbol{x}_n)} - \frac{1-y_n}{1-f(\boldsymbol{x}_n)}\right)\left(\frac{e^{-\boldsymbol{xw}}}{(1+e^{-\boldsymbol{xw}})^2}\right)\boldsymbol{x}_n = 0$$

$$\frac{\partial L}{\partial f(\boldsymbol{\theta})} \qquad \frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \qquad \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{w}}$$

Optimal LR parameters (1) – PROOF (e)

$$\sum_{n=1}^{N} - \left(\frac{y_n}{f(\boldsymbol{x}_n)} - \frac{1-y_n}{1-f(\boldsymbol{x}_n)}\right)\left(\frac{e^{-\boldsymbol{xw}}}{(1+e^{-\boldsymbol{xw}})^2}\right)\boldsymbol{x}_n = 0$$

expanding each of the two terms and substituting for $f(\boldsymbol{x}_n) = \frac{1}{1+e^{-\boldsymbol{xw}}}$

$$\sum_{n=1}^{N} - \left(\frac{y_n - f(\boldsymbol{x}_n)}{f(\boldsymbol{x}_n)\big(1-f(\boldsymbol{x}_n)\big)}\right)\Big(\boldsymbol{x}_n f(\boldsymbol{x}_n)\big(1-f(\boldsymbol{x}_n)\big)\Big) = 0$$

cancelling terms common to both numerator and denominator

$$\textstyle\sum_{n=1}^{N} - \boldsymbol{x}_n\big(y_n - f(\boldsymbol{x}_n)\big) = 0$$

dividing through by -1

$$\sum_{n=1}^{N} \boldsymbol{x}_n(f(\boldsymbol{x}_n) - y_n) = 0$$