Lecture 2

Elements of Machine Learning

Eg. want to predict whether a face is male / female, smiling, old / young, etc.

Key: Different face types have different models

model = mathematical description of the data

① Collect data — eg. photographs w/ faces.

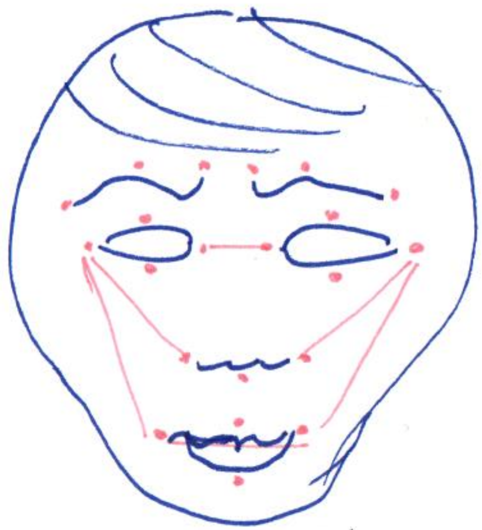
② preprocessing = changing data to simlify subsequent operations w/o losing relevant ~~ta~~ information

e.g. crop images to only contain <u>one</u> face,

center face

resize so all faces are same size.

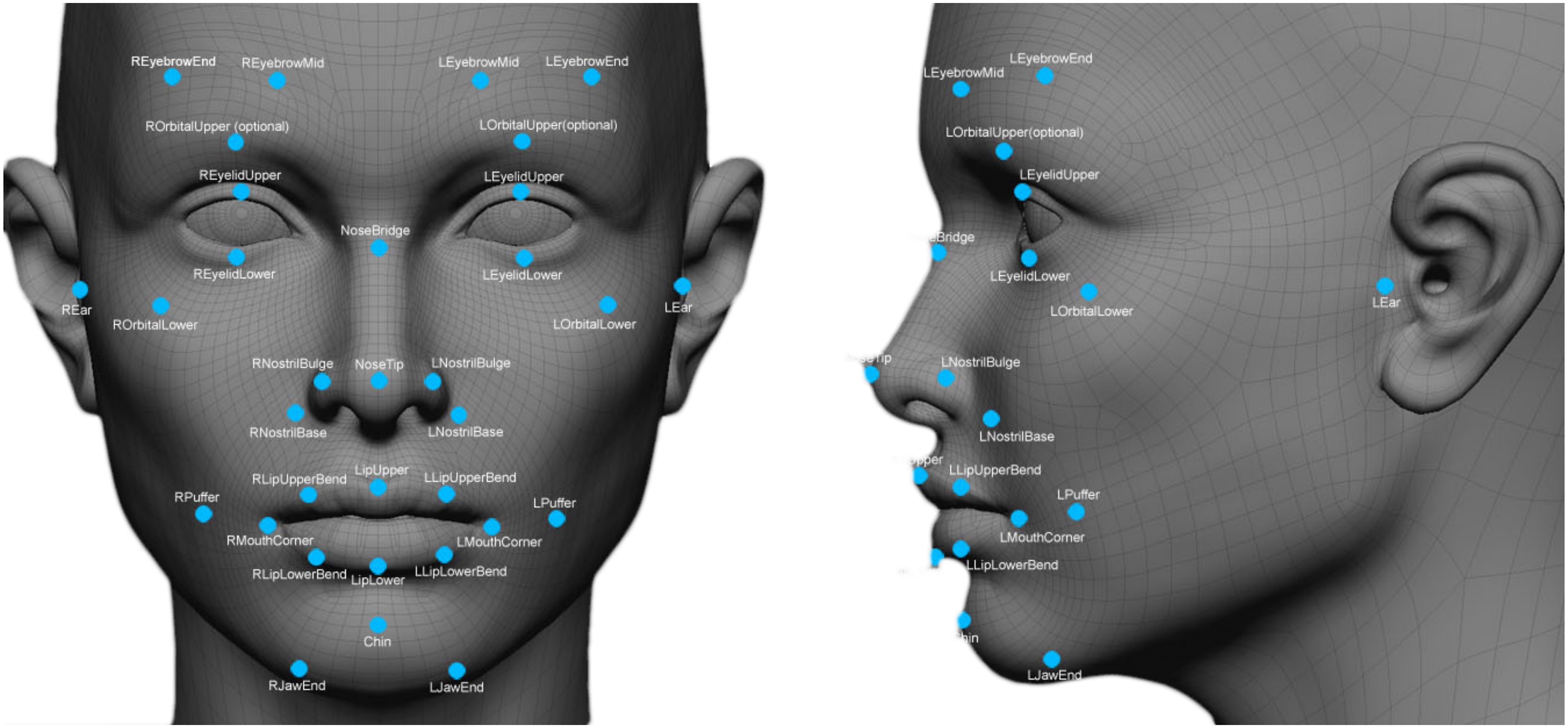
3 feature extraction = reduce raw data by extracting feature or properties relevant to model



e.g. distance between facial landmarks

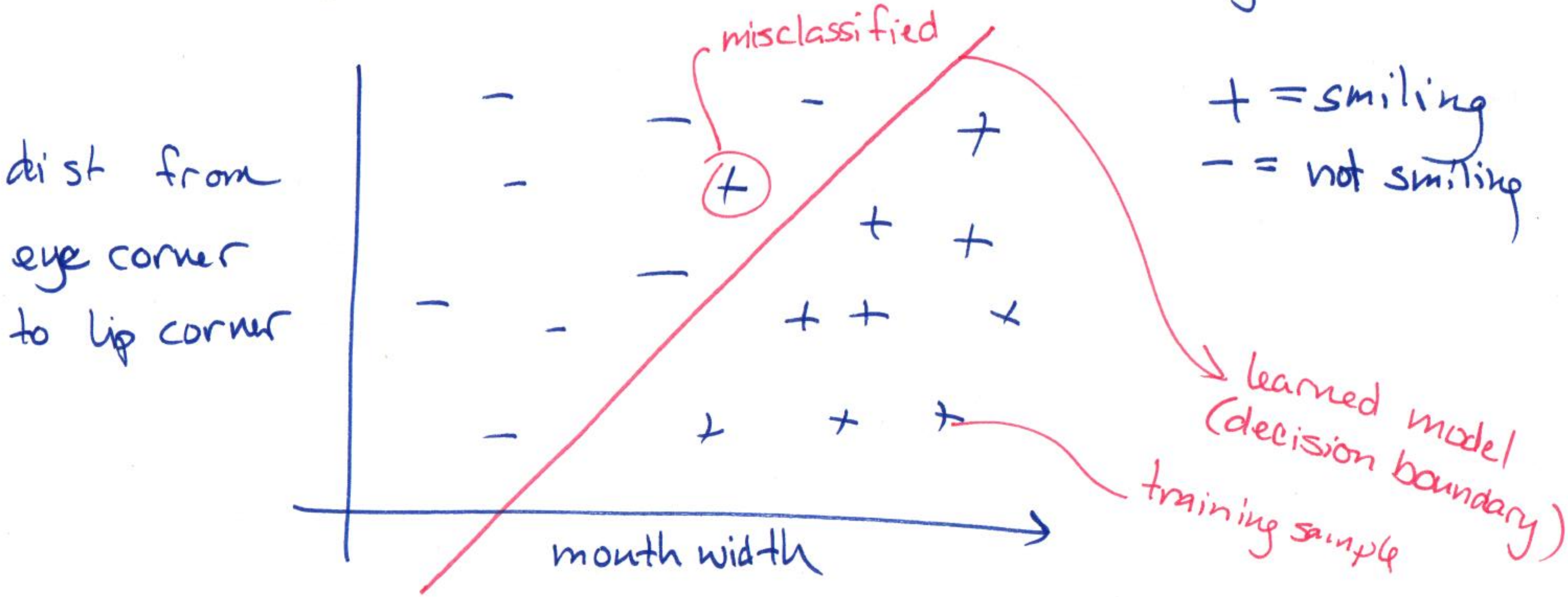
4 generate training samples = large collection of examples we can use to learn model

e.g. preprocess and extract features from many different images of faces

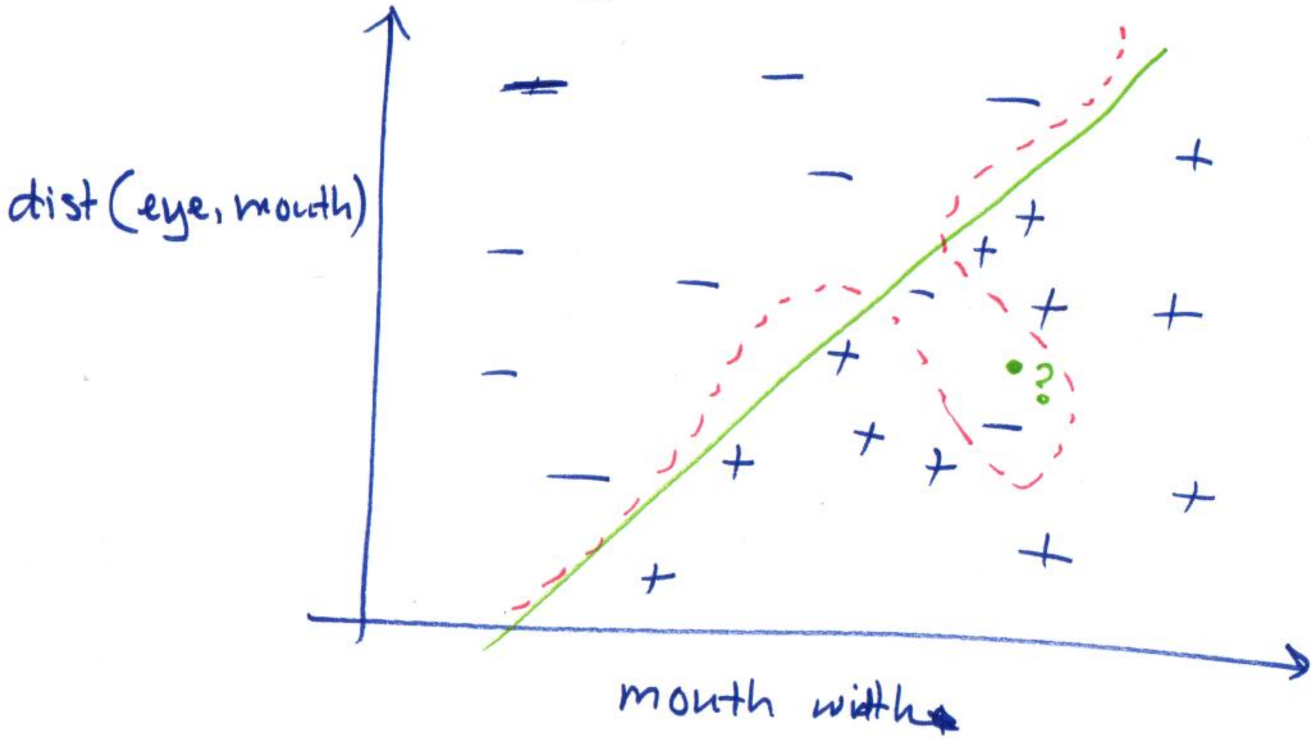

REyebrowEnd
REyebrowMid
LEyebrowMid
LEyebrowEnd
ROrbitalUpper (optional)
LOrbitalUpper(optional)
REyelidUpper
LEyelidUpper
NoseBridge
REyelidLower
LEyelidLower
REar
ROrbitalLower
LOrbitalLower
LEar
RNostrilBulge
NoseTip
LNostrilBulge
RNostrilBase
LNostrilBase
RLipUpperBend
LipUpper
LLipUpperBend
RPuffer
LPuffer
RMouthCorner
LMouthCorner
RLipLowerBend
LipLower
LLipLowerBend
Chin
RJawEnd
LJawEnd
LEyebrowMid
LEyebrowEnd
LOrbitalUpper(optional)
LEyelidUpper
LEyelidLower
LOrbitalLower
LEar
LNostrilBulge
LNostrilBase
LLipUpperBend
LPuffer
LMouthCorner
LLipLowerBend
LJawEnd

5 To learn model, we choose a loss function
= a measure of how well a model fits data
(e.g. % of samples misclassify "smiling")

4 finally learn the model = search over collection of candidate models or model parameters to find one that minimizes loss on training data



⑦ Characterize $\boxed{\text{generalization error}}$ = error of our predictions on new data that was not used for training



example model:

label $\hat{y} = w_1 x_1 + w_2 x_2$

weights   feature

This corresponds to an $\boxed{\text{inner product}}$

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \underline{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

(transpose:

$$\underline{x}^T = [x_1 \quad x_2]$$

inner product $= \langle \underline{x}, \underline{w} \rangle = \underline{x}^T \underline{w} = x_1 w_1 + x_2 w_2$

our model: $\hat{y} = \langle \underline{x}, \underline{w} \rangle$

if we have p features and p weights, then

$\langle \underline{x}, \underline{w} \rangle = \sum_{j=1}^{p} x_j w_j$, write $\underline{x}, \underline{w} \in \mathbb{R}^p$ (inside space of real-valued vecs. of length p)

now let's say we have n training samples $\underline{x}_i \in \mathbb{R}^p$,
$i = 1, 2, \dots, n$

our model says

$$\hat{y}_i = \langle \underline{x}_i , \underline{w} \rangle \quad \text{for } i = 1, 2, \dots, n$$
$$= \underline{x}_i^T \underline{w}$$

we can combine all these n equations:

$$\hat{\underline{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \underbrace{\begin{bmatrix} \text{---}\ \underline{x}_1^T\ \text{---} \\ \text{---}\ \underline{x}_2^T\ \text{---} \\ \vdots \\ \text{---}\ \underline{x}_n^T\ \text{---} \end{bmatrix}}_{\underbrace{X}_{\text{matrix}} \in \mathbb{R}^{n \times p} \text{ (n rows, p columns)}} \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix}}_{\underbrace{\underline{w}}_{\text{vector}} \in \mathbb{R}^p} \iff \hat{\underline{y}} = \underbrace{X \underline{w}}_{\text{matrix-vector multiplication}}$$

model $\hat{y} = X \underline{w}$

computing $X\underline{w}$ means taking the inner product of each row of X with w, and soring the results in a vector $\hat{y}$

note:

- number of columns in X = length of $\underline{w}$
- number of rows in X = length of $\hat{\underline{y}}$

Example:

$$X = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 3 \end{bmatrix}$$

2 features

3 training samples

1^{st} sample $= \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$$\underline{w} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

$$\underline{x}^T \underline{w} = x_1 w_1 + x_2 w_2$$

$$X \underline{w} = \begin{bmatrix} 2 \\ 4 \\ 12 \end{bmatrix}$$

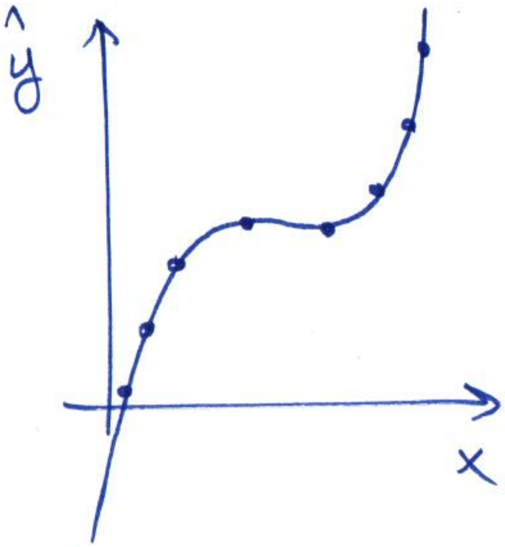
another perspective:

$X\underline{w}$ gives a weighted sum of the columns of X, where $\underline{w}$ are the weights

$$Xw = 2 \cdot \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} + 4 \cdot \begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 12 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 12 \end{bmatrix}$$

example: $(p = 1)$

model: $\hat{y}_i = w_3 \cdot x_i^3 + w_2 \cdot x_i^2 + w_1 \cdot x_i^1 + \underbrace{w_0 x_i^0}_{\equiv w_0}$



$$\underline{\hat{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \underbrace{\begin{bmatrix} x_1^3 & x_1^2 & x_1 & 1 \\ x_2^3 & x_2^2 & x_2 & 1 \\ \vdots & & & \\ x_n^3 & x_n^2 & x_n & 1 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} w_3 \\ w_2 \\ w_1 \\ w_0 \end{bmatrix}}_{\underline{w}}$$

"Vandermonde matrix"