CS 540 Introduction to Artificial Intelligence

Classification - KNN and Naive Bayes

Yudong Chen
University of Wisconsin-Madison

[Oct 14, 2021]

Slides created by Sharon Li [modified by Yudong Chen]

Announcement

Homework: HW5 due next Tuesday

Thursday, Sept 30	Machine Learning: Introduction
Tuesday, Oct 5	Machine Learning: Unsupervised Learning I
Thursday, Oct 7	Machine Learning: Unsupervised Learning II
Tuesday, Oct 12	Machine Learning: Linear regression
Thursday, Oct 14	Machine Learning: K - Nearest Neighbors & Naive Bayes

We will continue on supervised learning today

Today's outline

- K-Nearest Neighbors
- Maximum likelihood estimation
- Naive Bayes

Part I: K-nearest neighbors


Article | Talk

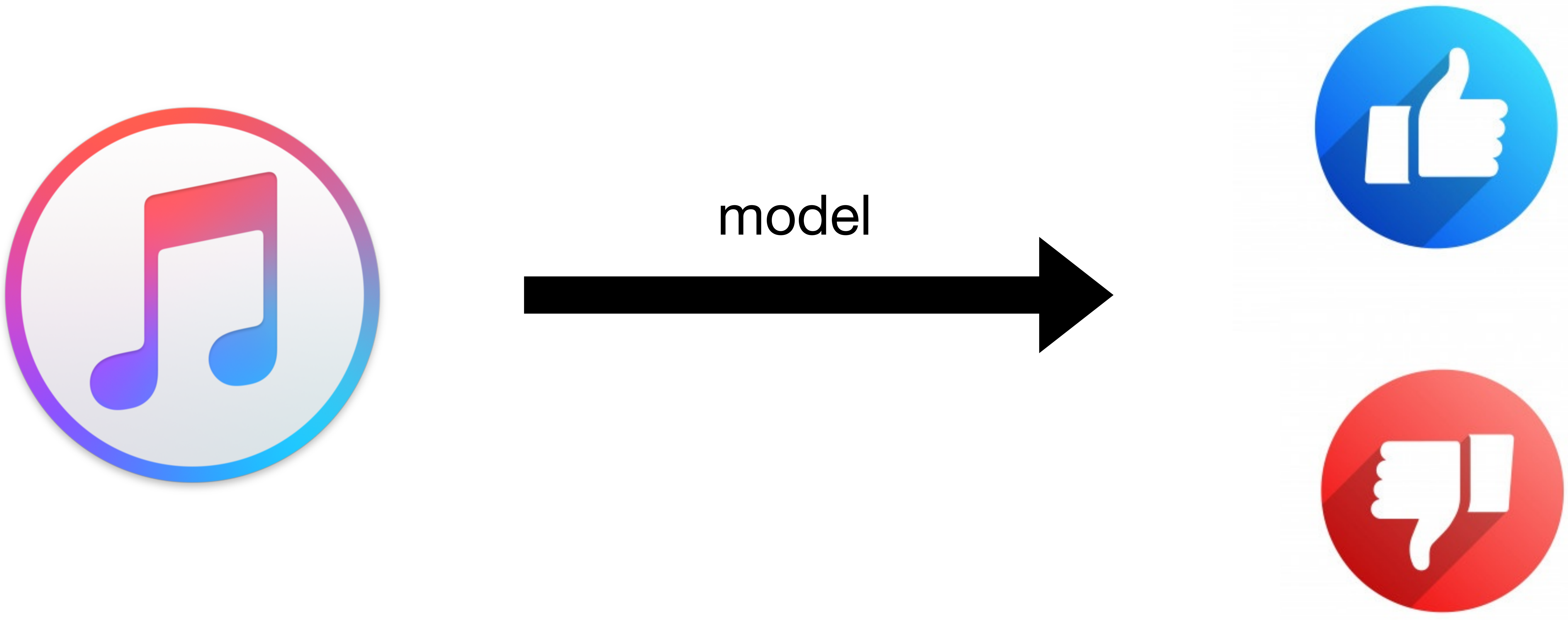
k-nearest neighbors algorithm

From Wikipedia, the free encyclopedia

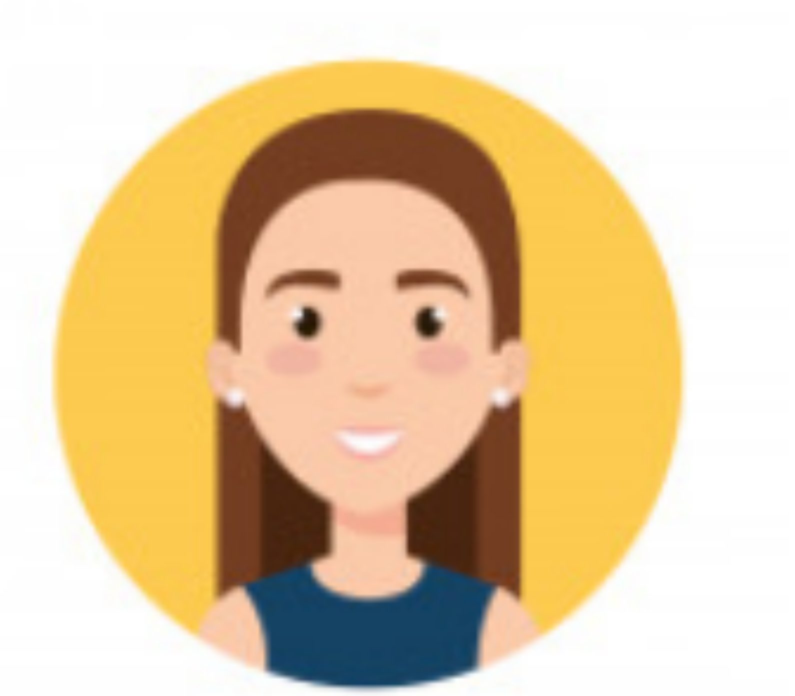
Not to be confused with k-means clustering.

(source: wiki)

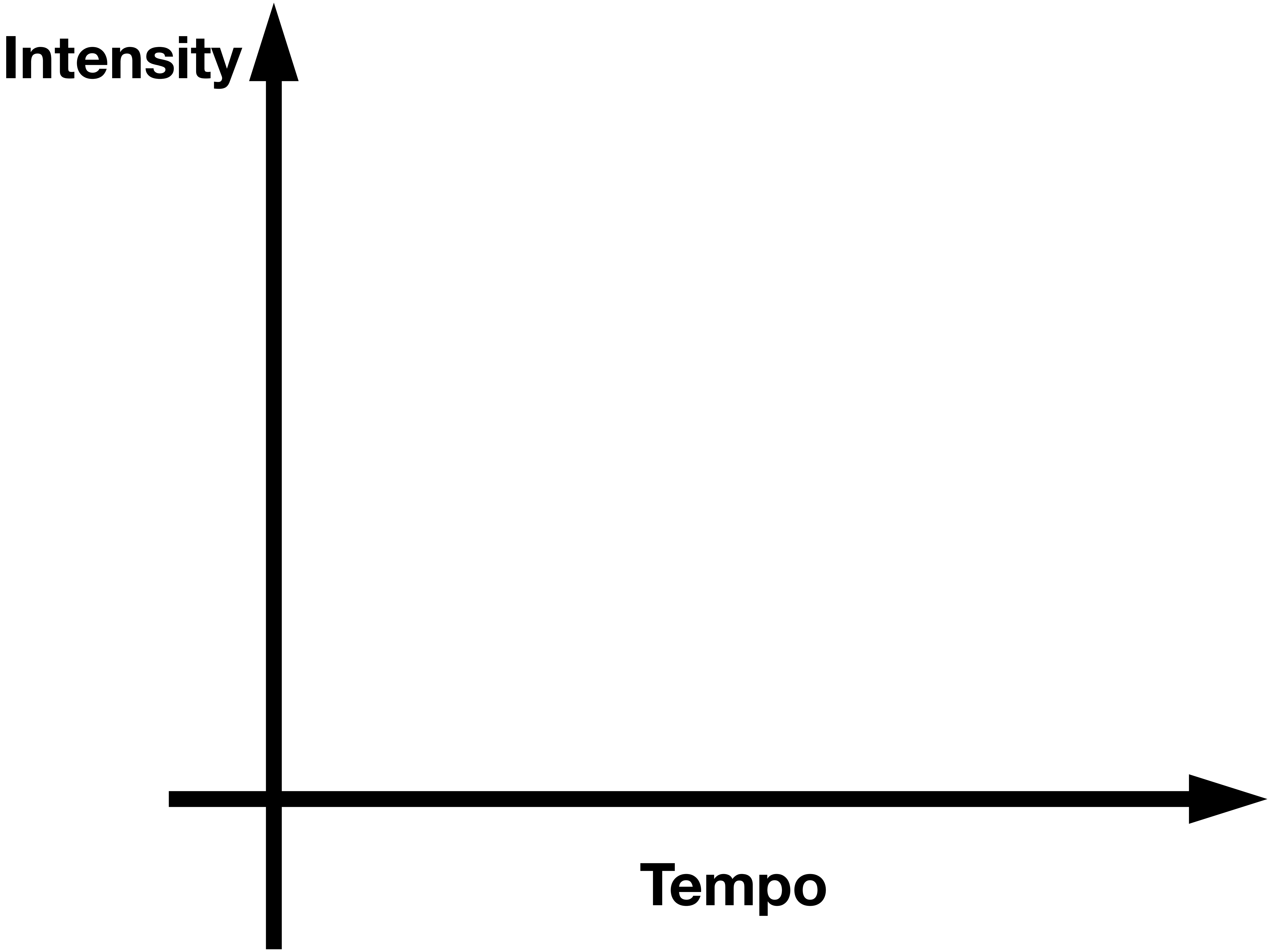
Example 1: Predict whether a user likes a song or not



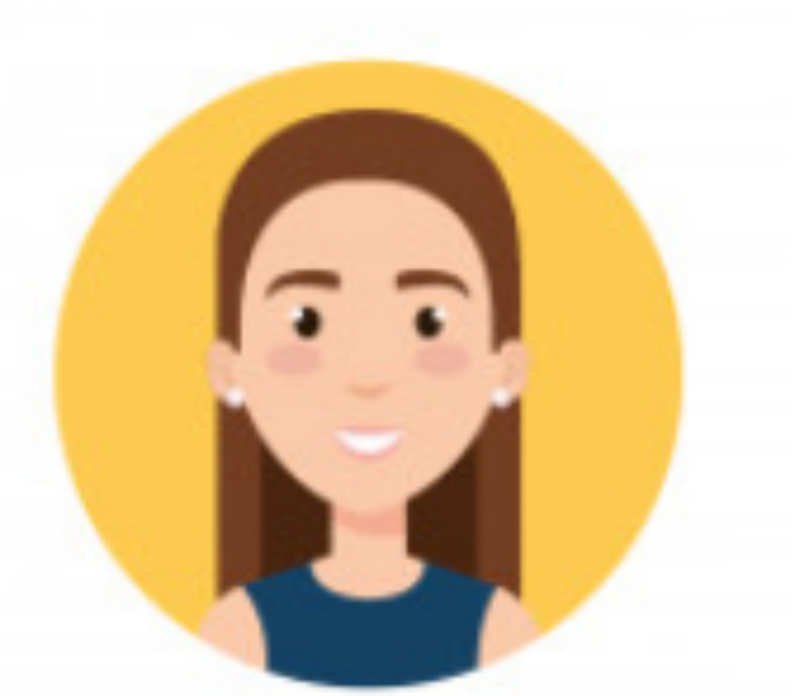
Example 1: Predict whether a user likes a song or not



User Sharon



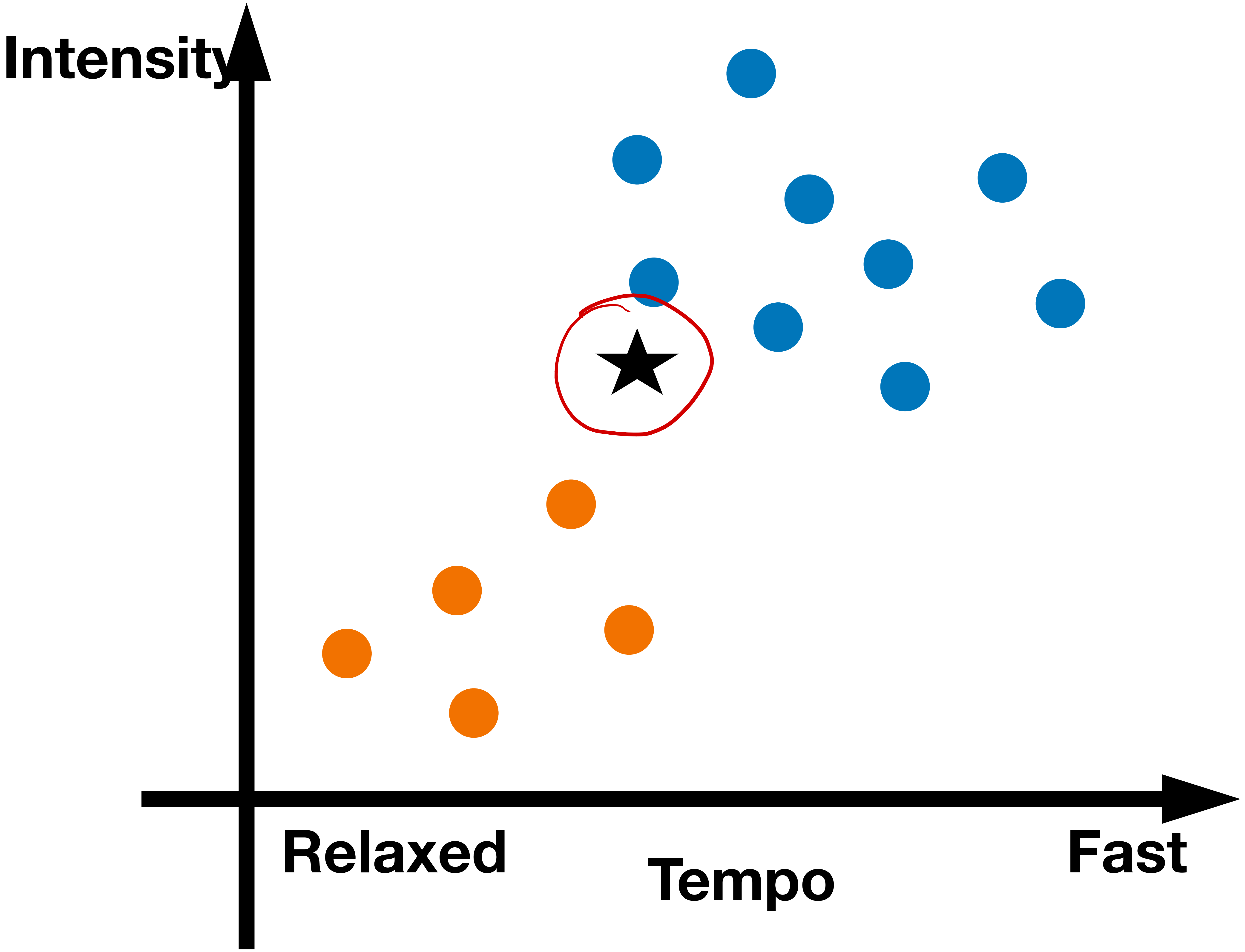
Example 1: Predict whether a user likes a song or not
1-NN



User Sharon

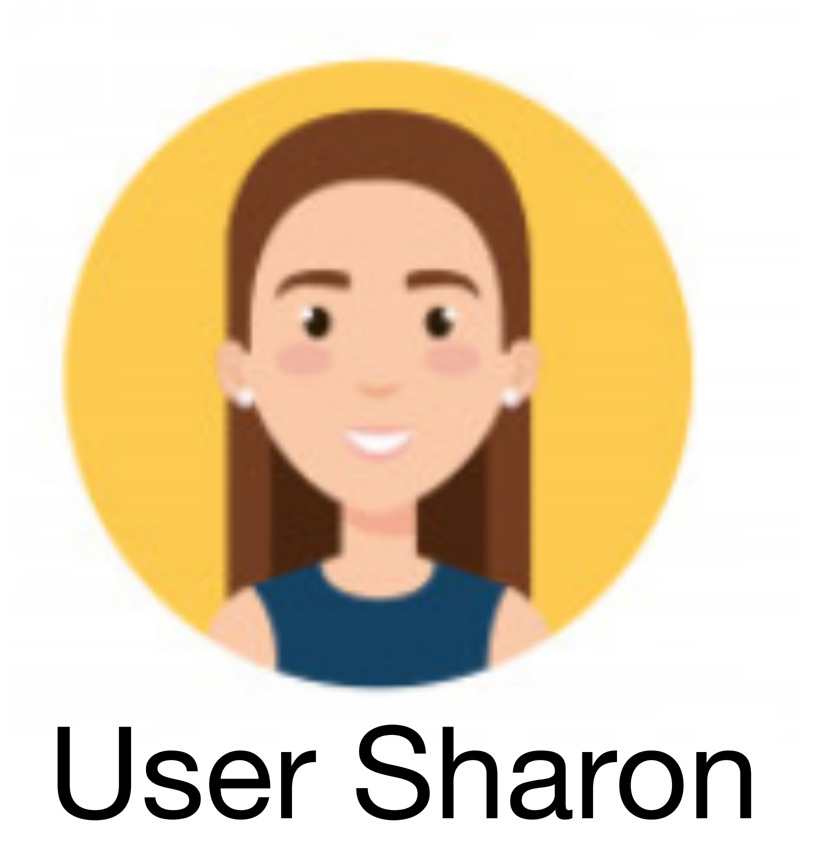
DisLike

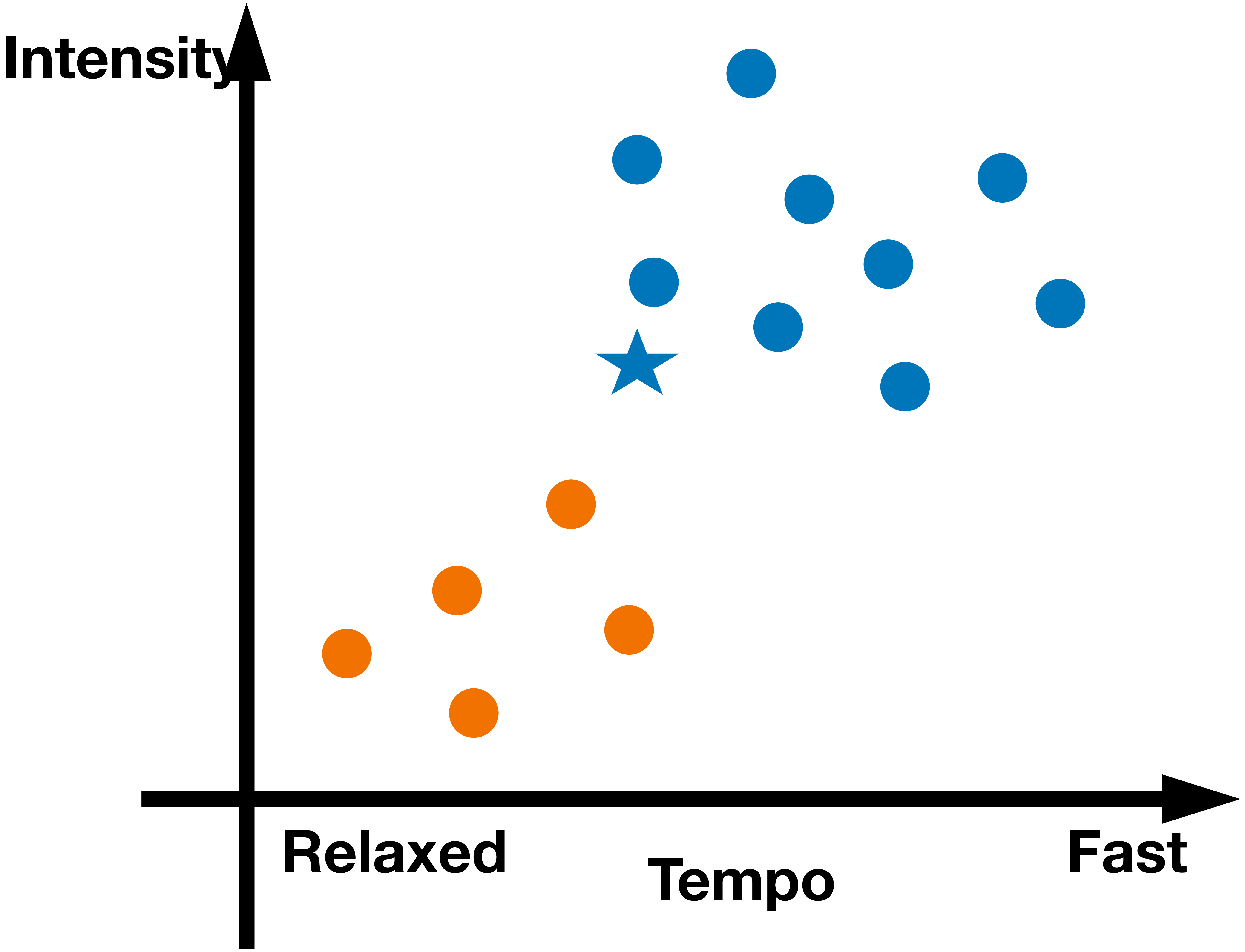
Like



Example 1: Predict whether a user likes a song or not

1-NN





K-nearest neighbors for classification

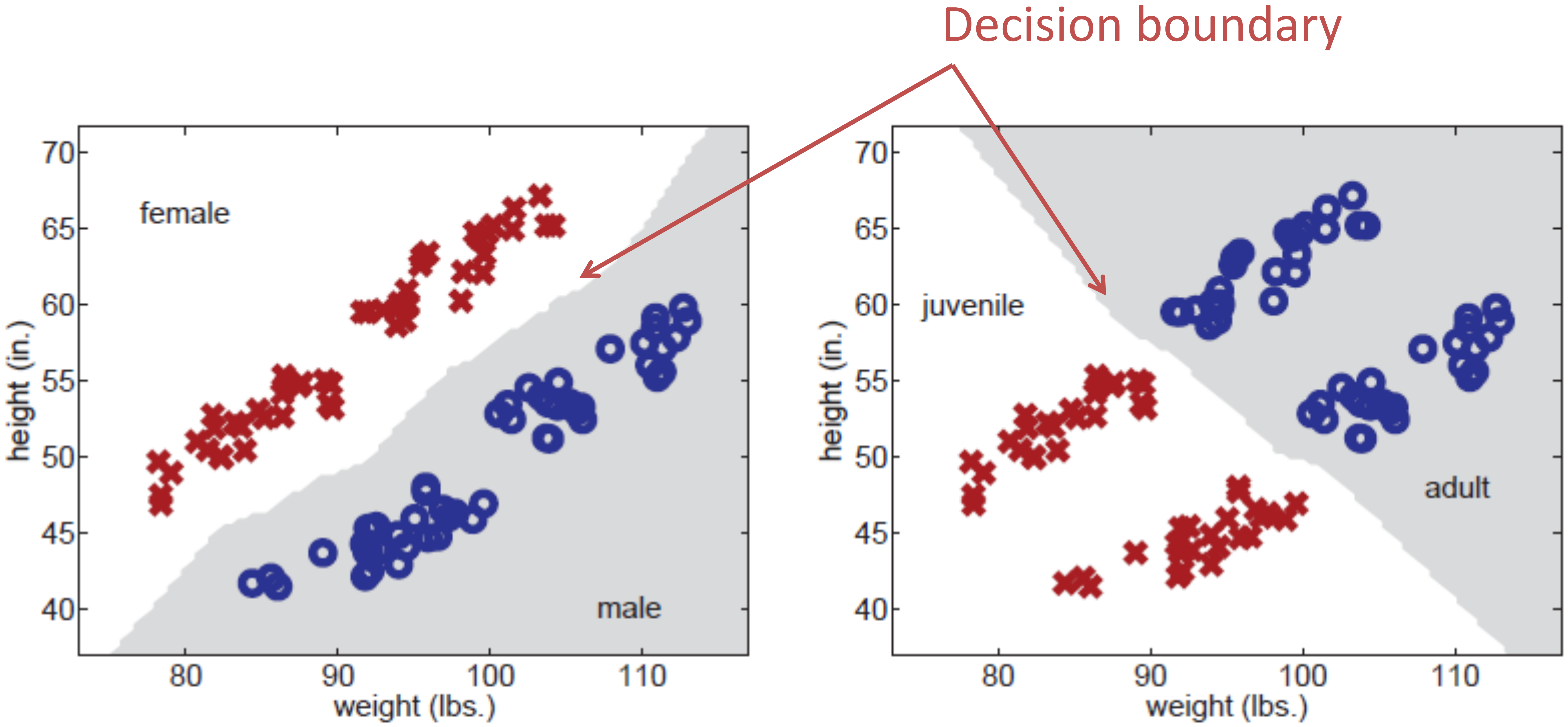
- **Input:** Training data $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$

  Distance function $d(\mathbf{x}_i, \mathbf{x}_j)$; number of neighbors k; test data $\mathbf{x}^*$

1. Find the k training instances $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}$ closest to $\mathbf{x}^*$ under $d(\mathbf{x}_i, \mathbf{x}_j)$
2. Output y^* as the majority class of $y_{i_1}, \ldots, y_{i_k}$. Break ties randomly.

Example 2: 1-NN for little green man

- Predict gender (M,F) from weight, height
- Predict age (adult, juvenile) from weight, height



(a) classification by gender

(b) classification by age

K-NN for regression

- What if we want regression?

- Instead of majority vote, take average of neighbors' labels
  - Given test point $\mathbf{x}^*$, find its k nearest neighbors $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}$
  - Output the predicted label $\frac{1}{k}(y_{i_1} + \ldots + y_{i_k})$

How can we determine distance?

suppose all features are discrete

- Hamming distance: count the number of features for which two instances differ

How can we determine distance?

$d(\vec{p}, \vec{q}) = \#\{i : p_i \neq q_i\}$

discrete

suppose all features are discrete

- Hamming distance: count the number of features for which two instances differ

$p = (p_1, p_2, p_3)$

$q = (q_1, q_2, q_3)$

$d(p,q) = \mathbb{1}\{p_1 \neq q_1\} + \sqrt{(p_2 - q_2)^2 + (p_3 - q_3)^2}$

suppose all features are continuous

continuous

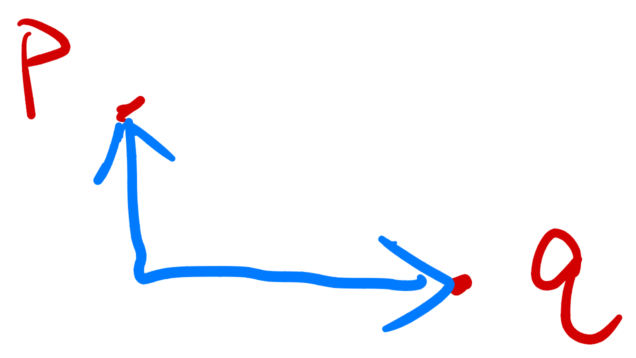
- Euclidean distance: sum of squared differences

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} = \|p - q\|_2$$

ℓ_2-distance.

- Manhattan distance: city block distance, ℓ_1-distance

$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |p_i - q_i|$$



How to pick the number of neighbors

k.

- Split data into training and **tuning sets** / validation set.
- Classify tuning set with different k
- Pick k that produces least tuning-set error

training set	validation set	test set
80	10	10
60	20	20

training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$

Validation set. $\{(x_{n+1}, y_{n+1}), \ldots, (x_{n+m}, y_{n+m})\}$

$k=1$ → decision boundary. → classify → validation error $(k=1)$

$k=2$ → validation error $(k=2)$

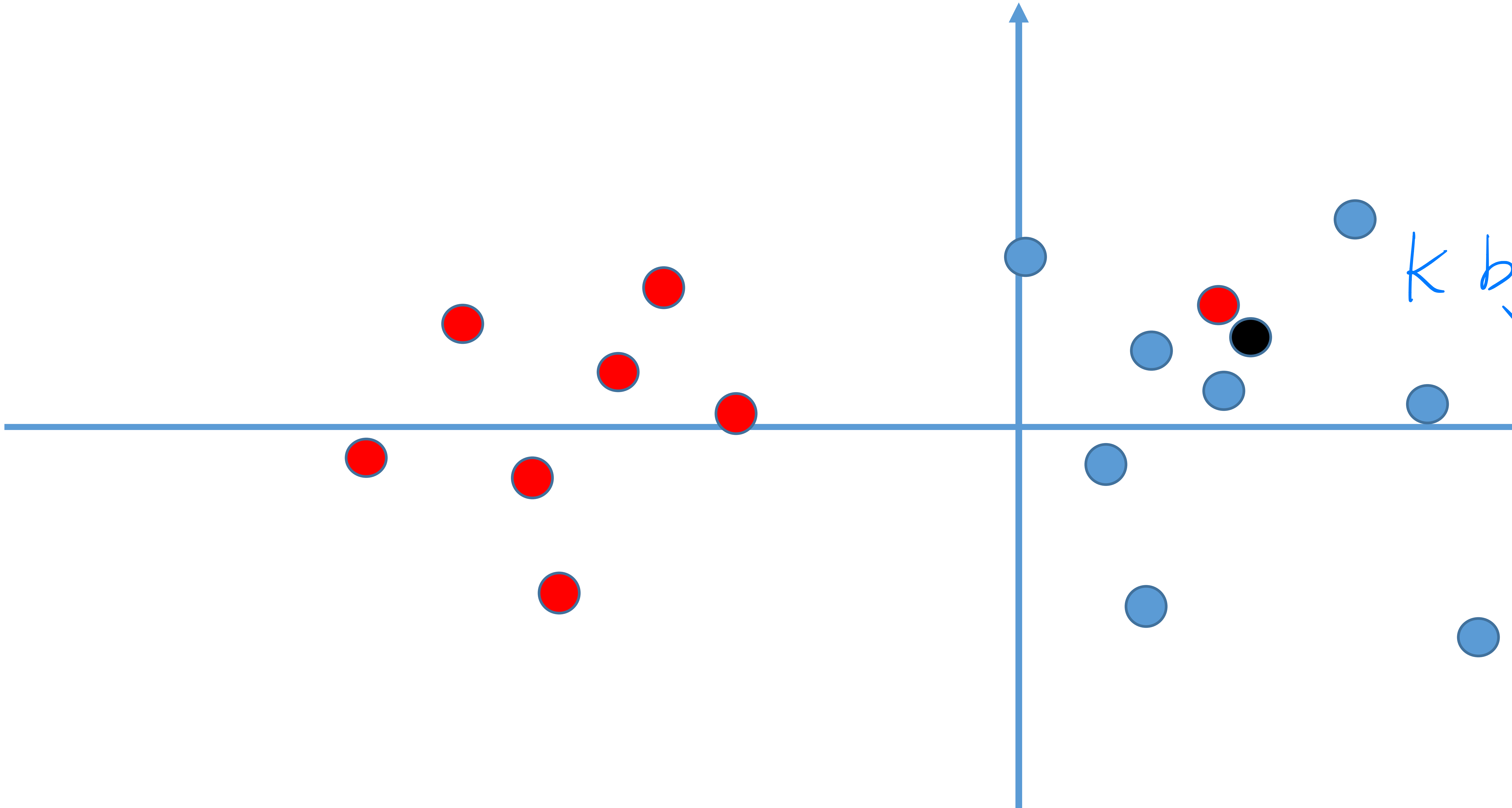
$\vdots$

$k=100$ → val-error $(k=100)$

Choose k w/ smallest validation error.

Effect of k

k small: sensitive
over-fit.
non-smooth bnd



k big: insensitive
~~over~~ underfit.
too smooth
bnd.

k = 20

What's the predicted label for the black dot using 1 neighbor? 3 neighbors?

Quiz break

Q1-1: K-NN algorithms can be used for:

- A Only classification
- B Only regression
- C Both

Quiz break

Q1-1: K-NN algorithms can be used for:

- A Only classification
- B Only regression
- C Both

Quiz break

Q1-2: Which of the following distance measure do we use in case of categorical variables in k-NN?

- A Hamming distance
- B Euclidean distance
- C Manhattan distance

Quiz break

Q1-2: Which of the following distance measure do we use in case
categorical variables in k-NN?

- A Hamming distance
- B Euclidean distance
- C Manhattan distance

Quiz break

Q1-3: Consider binary classification in 2D where the intended label of a point x = (x1, x2) is positive if x1>x2 and negative otherwise. Let the training set be all points of the form x = [4a, 3b] where a,b are integers. Each training item has the correct label that follows the rule above. With a 1NN classifier (Euclidean distance), which ones of the following points are labeled positive? Multiple answers. (4a, 3b)

- [5.52, 2.41] (4, 3) y=+
- [8.47, 5.84]
- [7,8.17]
- [6.7,8.88]

Quiz break

Q1-3: Consider binary classification in 2D where the intended label of a point x = (x1, x2) is positive if x1>x2 and negative otherwise. Let the training set be all points of the form x = [4a, 3b] where a,b are integers. Each training item has the correct label that follows the rule above. With a 1NN classifier (Euclidean distance), which ones of the following points are labeled positive? Multiple answers.

- [5.52, 2.41]
- [8.47, 5.84]
- [7,8.17]
- [6.7,8.88]

Nearest neighbors are
[4,3] => positive
[8,6] => positive
[8,9] => negative
[8,9] => negative
Individually.

Part II: Maximum Likelihood Estimation

Supervised Machine Learning

$f(x)$ = vote/average of K-nearest neighbor.

$\hat{y} = f_\theta(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d$.

Non-parametric (e.g., KNN)

vs.

Parametric

MLE

Supervised Machine Learning

Statistical modeling approach

Labeled training data (n examples)

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$$

drawn **independently** from
a fixed underlying distribution
(also called the i.i.d. assumption)

$(x_i, y_i) \overset{iid}{\sim}$ unknown probability model.

Supervised Machine Learning

Statistical modeling approach

Labeled training data (n examples) → **Learning algorithm** → **Classifier** $\hat{f}$

$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$

drawn **independently** from
a fixed underlying distribution
(also called the i.i.d. assumption)

select $\hat{f}(\theta)$ from a pool of probability models $\mathcal{F}$
that **best describe the data observed**

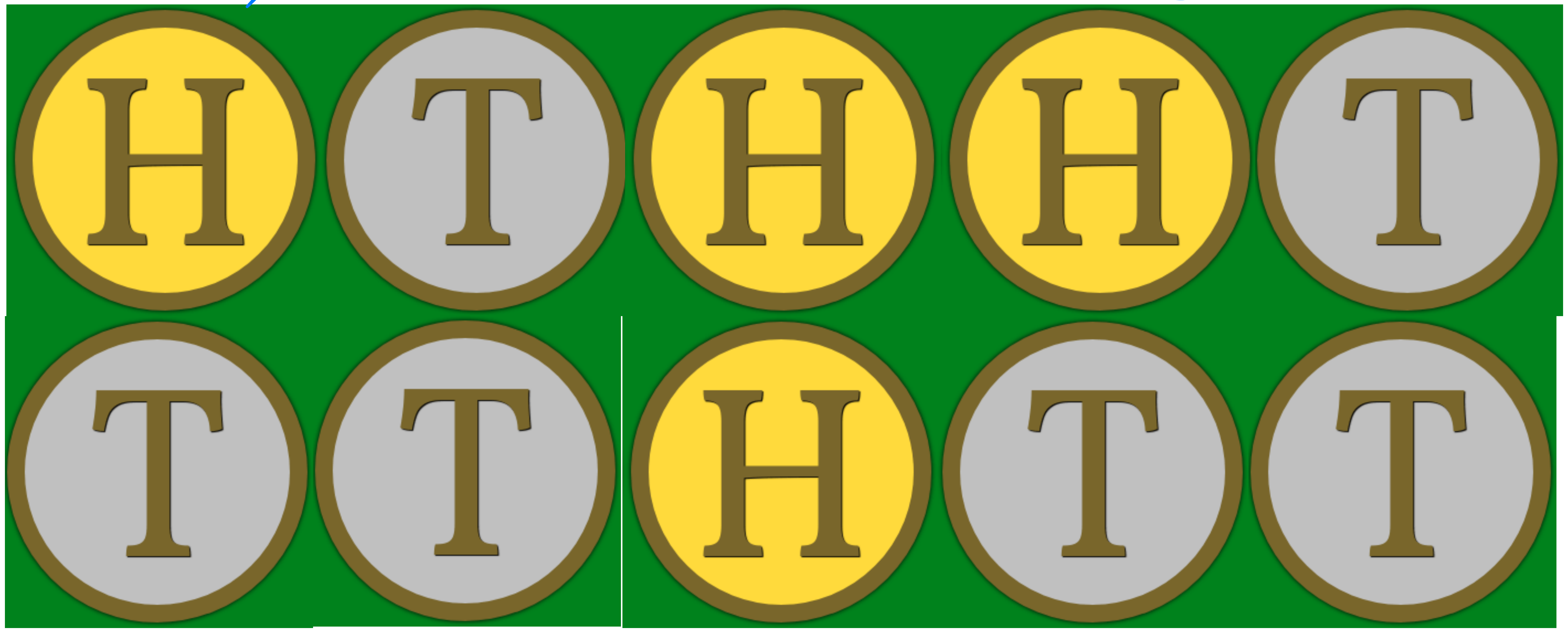
How to select $\hat{f} \in \mathscr{F}$?

- **Maximum likelihood (best fits the data)**
- Maximum a posteriori (best fits the data but incorporates prior assumptions)
- Optimization of ‘loss’ criterion (best discriminates the labels)

Maximum Likelihood Estimation: An Example

Flip a coin 10 times, how can you estimate $\theta = p(\text{Head})$?

$x_1, \ldots, x_n \sim \text{Bernoulli}(\theta)$



Intuitively, $\theta = 4/10 = 0.4$

How good is θ?

$L(\theta) = P(x_1, \cdots, x_n | \theta)$

It depends on how likely it is to generate the observed data

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ (Let's forget about label for a second)

Likelihood function $L(\theta) = \Pi_i p(\mathbf{x}_i \,|\, \theta)$

Under i.i.d assumption

Interpretation: How **probable** (or how likely) is the data given the probabilistic model p_θ?

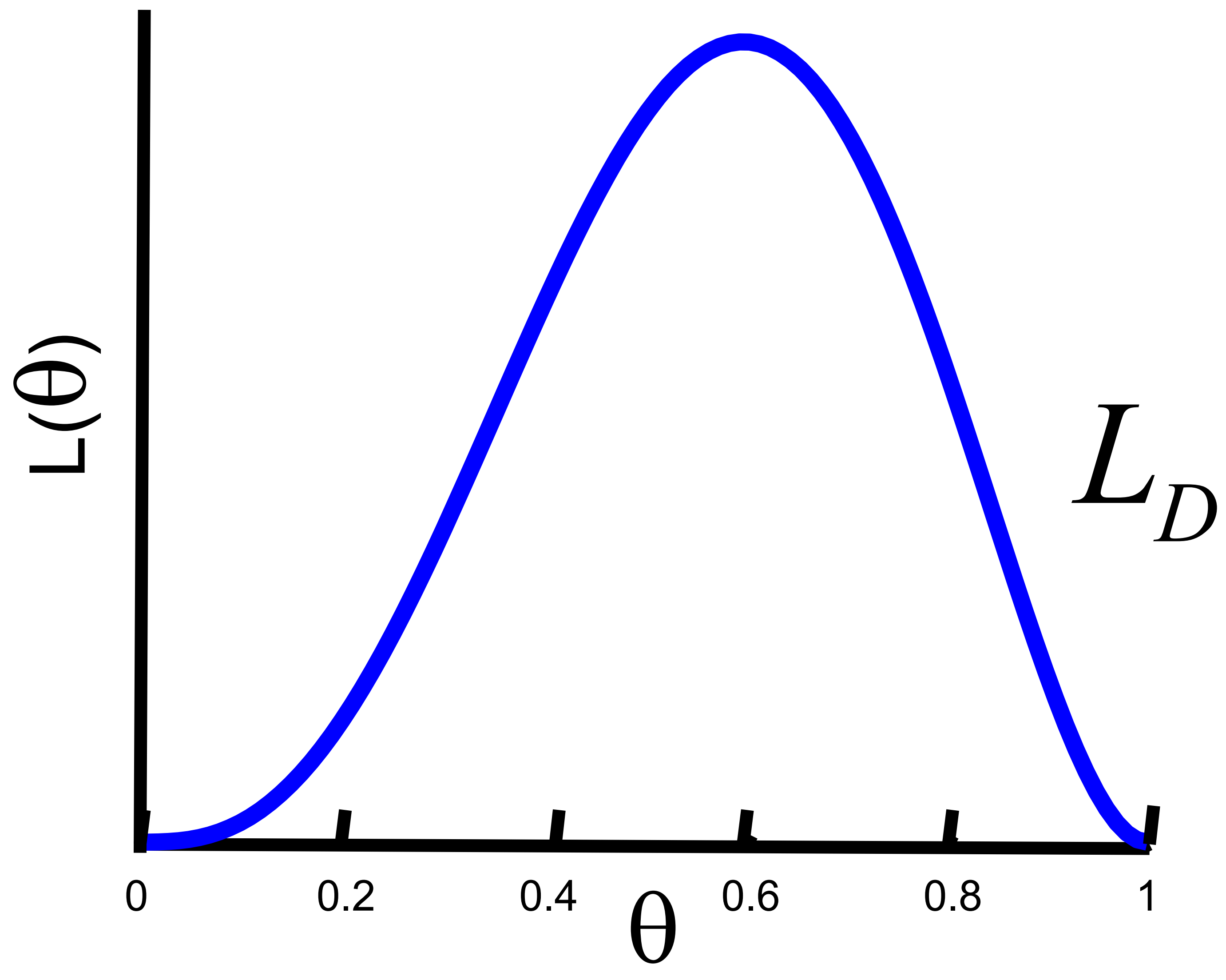
How good is θ?

It depends on how likely it is to generate the observed data

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ (Let's forget about label for a second)

Likelihood function $L(\theta) = \Pi_i p(\mathbf{x}_i \mid \theta)$

H,T, T, H, H



$$L_D(\theta) = \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta$$

Bernoulli distribution

Log-likelihood function

$$L_D(\theta) = \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta$$

$$= \theta^{N_H} \cdot (1-\theta)^{N_T}$$

N_H = # heads in data

N_T = # tails in data.

Log-likelihood function

$$\ell(\theta) = \log L(\theta)$$

$$= N_H \log \theta + N_T \log(1-\theta)$$

Maximum Likelihood Estimation (MLE)

Find optimal θ^* to maximize the likelihood function (and log-likelihood)

$$\theta^* = \arg\max N_H \log\theta + N_T \log(1-\theta)$$

$$\frac{\partial l(\theta)}{\partial\theta} = \frac{N_H}{\theta} - \frac{N_T}{1-\theta} = 0 \quad \Rightarrow \quad \theta^* = \frac{N_H}{N_T + N_H}$$

which confirms your intuition!

= frequency of heads.

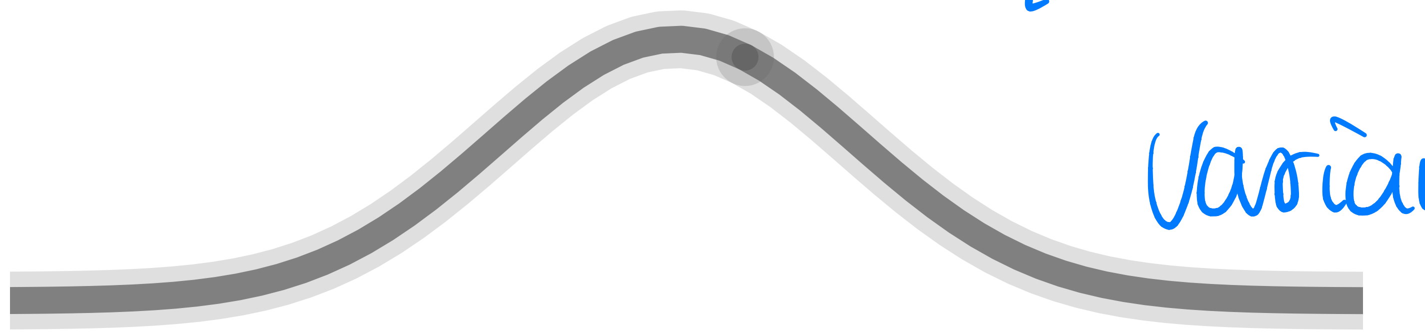
Maximum Likelihood Estimation: Gaussian Model

Fitting a model to heights of females

Observed some data (in inches): 60, 62, 53, 58,… $\in \mathbb{R}$

$$\{x_1, x_2, \ldots, x_n\}$$ ~ Gaussian w/ mean μ

variance σ^2.



Model class: Gaussian model

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

So, what's the MLE for the given data?

Estimating the parameters in a Gaussian

- **Mean**

$$\mu = \mathbf{E}[x] \text{ hence } \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Variance**

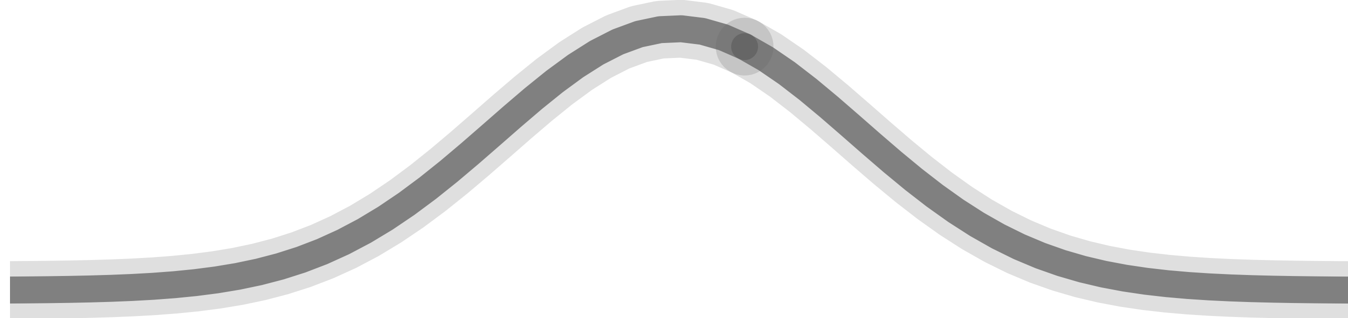
$$\sigma^2 = \mathbf{E}\left[(x-\mu)^2\right] \text{ hence } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Why?

Maximum Likelihood Estimation: Gaussian Model

Observe some data (in inches): $x_1, x_2, \ldots, x_n \in \mathbb{R}$

Assume that the data is drawn from a Gaussian



$$L(\mu, \sigma^2 \mid X) = \prod_{i=1}^{n} p(x_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Fitting parameters is maximizing likelihood w.r.t μ, σ^2
(maximize likelihood that data was generated by model)

MLE

$$\arg\max_{\mu, \sigma^2} \prod_{i=1}^{n} p(x_i; \mu, \sigma^2)$$

Maximum Likelihood

- Estimate parameters by finding ones that explain the data

$$\arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} p(x_i;\mu,\sigma^2) = \arg\min_{\mu,\sigma^2} \left[-\log \prod_{i=1}^{n} p(x_i;\mu,\sigma^2) \right]$$

negative log likelihood

- **Decompose likelihood**

$$\sum_{i=1}^{n} \left[\frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(x_i-\mu)^2 \right] = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$0 = \frac{\partial}{\partial \mu} = 0 + \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu)$$

Minimized for $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Maximum Likelihood

- Estimating the variance

$$-\ell(\mu, \sigma^2) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Maximum Likelihood

- Estimating the variance

$$\frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

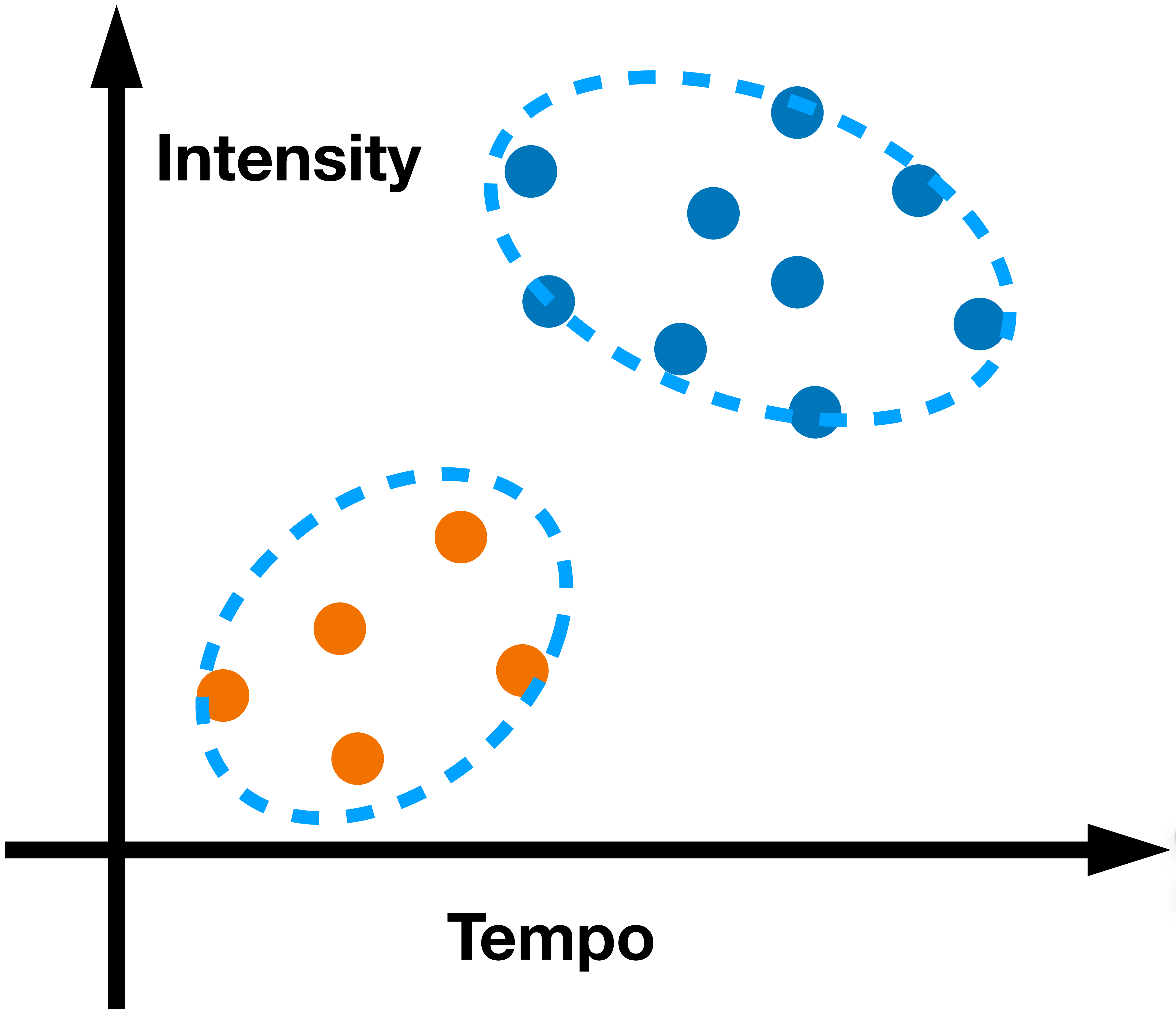
- Take derivatives with respect to it

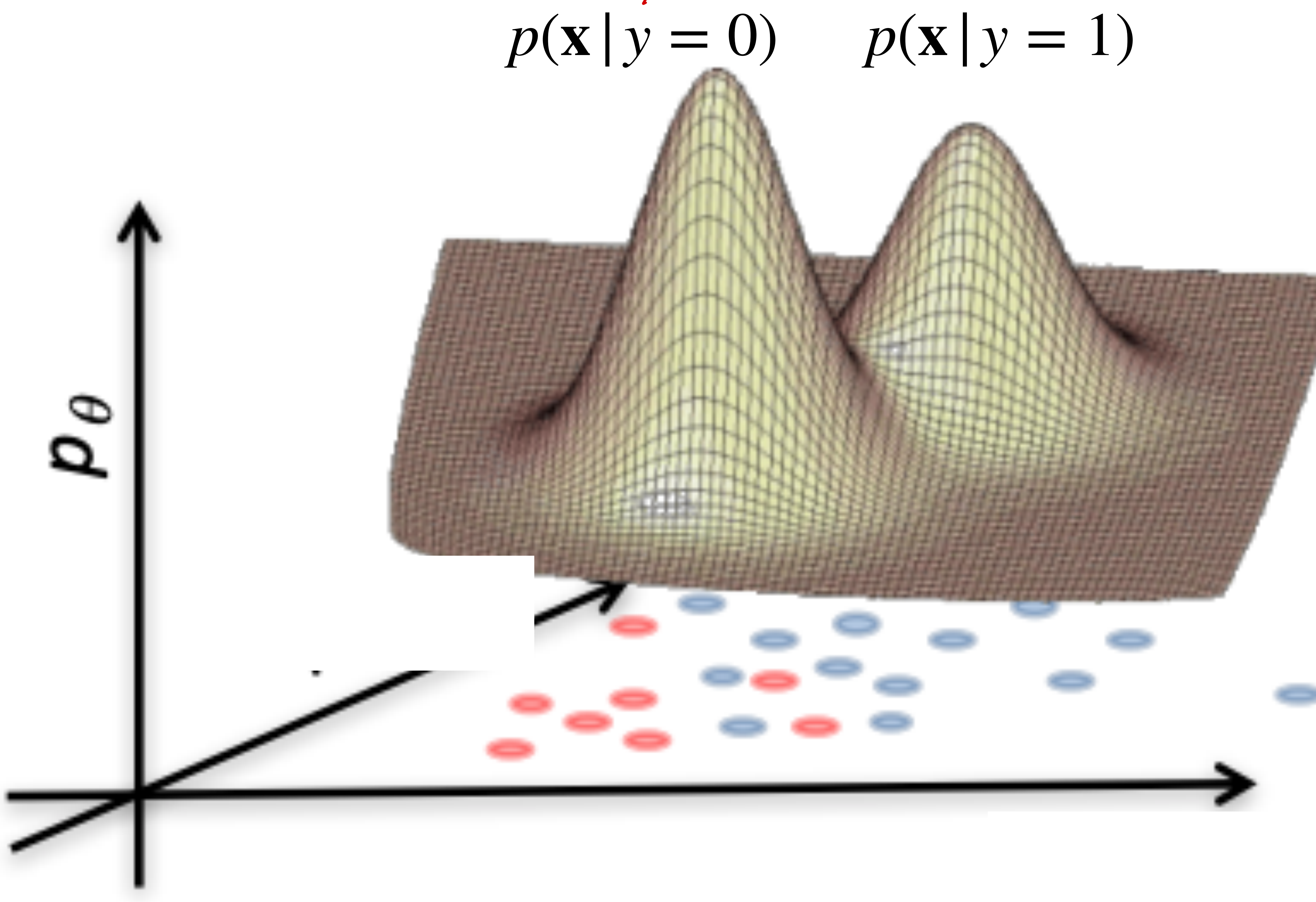
$$\partial_{\sigma^2}[\,\cdot\,] = \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

$$\implies \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Classification via MLE

class conditional probability





Classification via MLE

$$\hat{y} = \hat{f}(\mathbf{x}) = \arg\max_{y} p(y \mid \mathbf{x})$$

(Prediction)

(Posterior)

Classification via MLE

class conditional prob

$$\hat{y} = \hat{f}(\mathbf{x}) = \arg\max_y p(y \mid \mathbf{x})$$ (Posterior)

(Prediction)

prior probability.

$$= \arg\max_y \frac{p(\mathbf{x} \mid y) \cdot p(y)}{p(\mathbf{x})}$$ (by Bayes' rule)

$$= \arg\max_y \; p(\mathbf{x} \mid y) p(y)$$

Using labelled training data, learn **class priors** and **class conditionals**

Bayes Classifier.

Quiz break

Q2-2: True or False
Maximum likelihood estimation is the same regardless of whether we maximize the likelihood or log-likelihood function.

- A True
- B False

Quiz break

Q2-2: True or False
Maximum likelihood estimation is the same regardless of whether we maximize the likelihood or log-likelihood function.

- A True
- B False

log(·) is monotonically increasing.

Quiz break

Q2-3: Suppose the weights of randomly selected American female college students are normally distributed with unknown mean μ and standard deviation σ. A random sample of 10 American female college students yielded the following weights in pounds: 115 122 130 127 149 160 152 138 149 180. Find a maximum likelihood estimate of μ.

- A 132.2
- B 142.2
- C 152.2
- D 162.2

$$\frac{1}{10}(115+122+\cdots+180)$$

Quiz break

Q2-3: Suppose the weights of randomly selected American female college students are normally distributed with unknown mean μ and standard deviation σ. A random sample of 10 American female college students yielded the following weights in pounds: 115 122 130 127 149 160 152 138 149 180. Find a maximum likelihood estimate of μ.

- A 132.2
- B 142.2
- C 152.2
- D 162.2

Part II: Naïve Bayes

Bayesian classification

$$\hat{y} = \arg\max p(y \mid \mathbf{x})$$ (Posterior)

(Prediction)

$$= \arg\max \frac{p(\mathbf{x} \mid y) \cdot p(y)}{p(\mathbf{x})}$$ (by Bayes' rule)

$$= \arg\max \; p(\mathbf{x} \mid y) p(y)$$

Bayesian classification

What if **x** has multiple attributes $\mathbf{x} = \{X_1, \ldots, X_k\}$

$$\hat{y} = \arg\max_y p(y \mid X_1, \ldots, X_k)$$

(Prediction)

(Posterior)

Bayesian classification

What if **x** has multiple attributes $\mathbf{x} = \{X_1, \ldots, X_k\}$

$$\hat{y} = \arg\max_y p(y \mid X_1, \ldots, X_k) \quad \text{(Posterior)}$$

(Prediction)

$$= \arg\max_y \frac{p(X_1, \ldots, X_k \mid y) \cdot p(y)}{p(X_1, \ldots, X_k)} \quad \text{(by Bayes' rule)}$$

Independent of y

Bayesian classification

What if $\mathbf{x}$ has multiple attributes $\mathbf{x} = \{X_1, \ldots, X_k\}$

$$\hat{y} = \arg\max_y p(y \,|\, X_1, \ldots, X_k)$$ (Posterior)

(Prediction)

$$= \arg\max_y \frac{p(X_1, \ldots, X_k \,|\, y) \cdot p(y)}{p(X_1, \ldots, X_k)}$$ (by Bayes' rule)

$$= \arg\max_y \; p(X_1, \ldots, X_k \,|\, y) \; p(y)$$

Class conditional likelihood

Class prior

Naïve Bayes Assumption

Conditional independence of feature attributes

$$p(X_1, \ldots, X_k \mid y)p(y) = \Pi_{i=1}^{k} p(X_i \mid y)p(y)$$

Easier to estimate

(using MLE!)

$p(y) = p(\text{year})$

$p(y=2016)$

$p(y=2020)$

$p(w_i \mid y=2016)$

$p(w_i \mid y=2020)$

Example 1: Play outside or not?

- If weather is sunny, would you likely to play outside?

Posterior probability p(Yes | ) vs. p(No | )

Example 1: Play outside or not?

- If weather is sunny, would you likely to play outside?

Posterior probability p(Yes | ) vs. p(No | )

- Weather = {Sunny, Rainy, Overcast}
- Play = {Yes, No}
- Observed data {Weather, play on day *m*}, m={1,2,…,N}

Example 1: Play outside or not?

- If weather is sunny, would you likely to play outside?

Posterior probability p(Yes | ) vs. p(No | )

- Weather = {Sunny, Rainy, Overcast}
- Play = {Yes, No}
- Observed data {Weather, play on day *m*}, m={1,2,…,N}

p(Play | ) = p( | Play) p(Play) / p()

Bayes rule

Example 1: Play outside or not?

- **Step 1**: Convert the data to a frequency table of Weather and Play

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Example 1: Play outside or not?

Step 1: Convert the data to a frequency table of Weather and Play

Step 2: Based on the frequency table, calculate **likelihoods** and **priors**

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

$p(\text{Play} = \text{Yes}) = 0.64$

p( | Yes) = 3/9 = 0.33

https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Example 1: Play outside or not?

Step 3: Based on the likelihoods and priors, calculate posteriors

P(Yes| )

=P( |Yes)*P(Yes)/P()

?

P(No| )

=P( |No)*P(No)/P()

?

Example 1: Play outside or not?

Step 3: Based on the likelihoods and priors, calculate posteriors

P(Yes| ☀)
=P(☀ |Yes)*P(Yes)/P(☀)
=0.33*0.64/0.36
=0.6

P(No| ☀)
=P(☀ |No)*P(No)/P(☀)
=0.4*0.36/0.36
=0.4

P(Yes| ☀) > P(No| ☀) go outside and play!

Quiz break

Q3-1: Which of the following about Naive Bayes is incorrect?

Features

- A Attributes can be nominal or numeric
- B Attributes are equally important
- C Attributes are statistically dependent of one another given the class value
- D Attributes are statistically independent of one another given the class value
- E All of above

Quiz break

Q3-1: Which of the following about Naive Bayes is incorrect?

- A Attributes can be nominal or numeric
- B Attributes are equally important
- C Attributes are statistically dependent of one another given the class value
- D Attributes are statistically independent of one another given the class value
- E All of above

Quiz break

$y \in \{1, 2, \dots, 32\}$

Q3-2: Consider a [32-class] classification problem with two binary features, $x_1, x_2 \in \{0,1\}$. Suppose $P(Y = y) = 1/32$, $P(x_1 = 1 | Y = y) = y/46$, $P(x_2 = 1 | Y = y) = y/62$. Which class will naive Bayes classifier produce on a test item with $x_1 = 1$ and $x_2 = 0$?

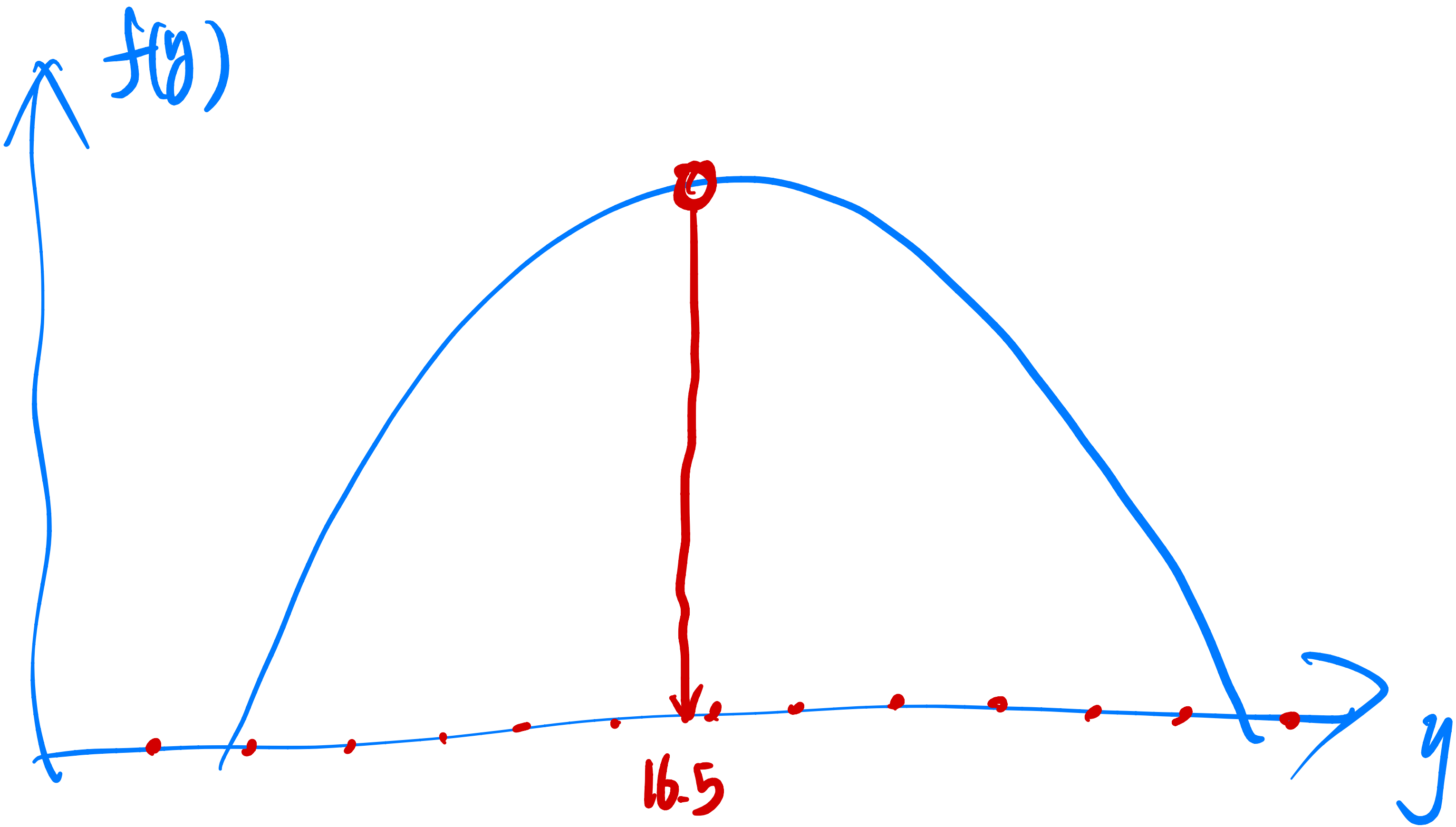
- A 16
- B 26
- C 31
- D 32

$$\arg\max_{y \in \{1,\dots,32\}} P(x_1=1, x_2=0 | Y=y) \cdot P(Y=y)$$

$$= \arg\max_y P(x_1=1|Y=y) P(x_2=0|Y=y) P(Y=y)$$

$$= \arg\max_y \frac{y}{46}\left(1-\frac{y}{62}\right)\cdot\frac{1}{32} = f(y)$$

$$f'(y) = \frac{1}{32}\left(\frac{1}{46} - \frac{2y}{46\cdot 62}\right) = 0 \Rightarrow y = 32$$


f(y)
16.5
y

Quiz break

Q3-2: Consider a classification problem with two binary features, $x_1, x_2 \in \{0,1\}$. Suppose $P(Y = y) = 1/32$, $P(x_1 = 1 | Y = y) = y/46$, $P(x_2 = 1 | Y = y) = y/62$. Which class will naive Bayes classifier produce on a test item with $x_1 = 1$ and $x_2 = 0$?

- A 16
- B 26
- C 31
- D 32

$$\arg\max_{y \in \{1,\dots,32\}} p(Y=y \mid x_1=1, x_2=0)$$

$$= \arg\max_y p(x_1=1, x_2=0 \mid Y=y)\, p(Y=y)$$

$$= \arg\max_y P(x_1=1 \mid Y=y)\, P(x_2=0 \mid Y=y)\, P(Y=y)$$

$$= \arg\max_y \frac{y}{46} \cdot \left(1 - \frac{y}{62}\right) \cdot \frac{1}{32}$$

$$\frac{\partial}{\partial y} = \frac{1}{46} - \frac{2}{46 \cdot 62} y = 0$$

$$\Rightarrow y = 31$$

Quiz break

Q3-3: Consider the following dataset showing the result whether a person has passed or failed the exam based on various factors. Suppose the factors are independent to each other. We want to classify a new instance with Confident=Yes, Studied=Yes, and Sick=No.

Confident	Studied	Sick	Result
Yes	No	No	Fail
Yes	No	Yes	Pass
No	Yes	Yes	Fail
No	Yes	No	Pass
Yes	Yes	Yes	Pass

- A Pass
- B Fail

Quiz break

Q3-3: Consider the following dataset showing the result whether a person has passed or failed the exam based on various factors. Suppose the factors are independent to each other. We want to classify a new instance with Confident=Yes, Studied=Yes, and Sick=No.

Confident	Studied	Sick	Result
Yes	No	No	Fail
Yes	No	Yes	Pass
No	Yes	Yes	Fail
No	Yes	No	Pass
Yes	Yes	Yes	Pass

- A Pass
- B Fail

What we’ve learned today…

- K-Nearest Neighbors
- Maximum likelihood estimation
  - Bernoulli model
  - Gaussian model
- Naive Bayes
  - Conditional independence assumption

Thanks!

Based on slides from Sharon Li, Xiaojin (Jerry) Zhu, Yingyu Liang and James McInerney