CS 540 Introduction to Artificial Intelligence

Classification - KNN and Naive Bayes

Sharon Yixuan Li

University of Wisconsin-Madison

March 2, 2021

Announcement

note @458 114 views

Actions

Midterm information

We are about 2.5 weeks away now; below you'll find useful information. We'll answer more questions on the format as we get closer.

The format of the midterm will include a mix of questions. There will be conceptual questions which have multiple choice or short sentence answers, but also computational questions where you'll be asked to perform a simple version of an algorithm, or related components, where you will show your work. The questions will vary from easy to hard.

Topics we'll cover include (but not strictly limited to)
- Probability: joint & conditional prob., inference, means and variances
- PCA: use and implementation
- NLP: language models, n-grams, evaluation
- General setup for ML: Supervised vs unsupervised, classification vs regression, loss functions, train vs test, overfitting
- Unsupervised learning: clustering (k-means & hierarchical), histograms, density estimation
- Linear models & linear regression
- kNN, naive Bayes, ML vs MAP, neural networks (in upcoming lectures)

Anything you did on the homeworks is fair game as well.

To help get you used to the types of questions being asked, we'll release a set of sample questions one week before (i.e., Weds. March 10th).
#pin

announcements

https://piazza.com/class/kk1k70vbawp3ts?cid=458

Announcement

Homework: HW4 review on Thursday / HW5 release today

Class roadmap

Tuesday, Feb 16	Machine Learning: Introduction
Thursday, Feb 18	Machine Learning: Unsupervised Learning I
Tuesday, Feb 23	Machine Learning: Unsupervised Learning II
Thursday, Feb 25	Machine Learning: Linear regression
Tuesday, March 2	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

WIKIPEDIA
The Free Encyclopedia

Main page

Article Talk

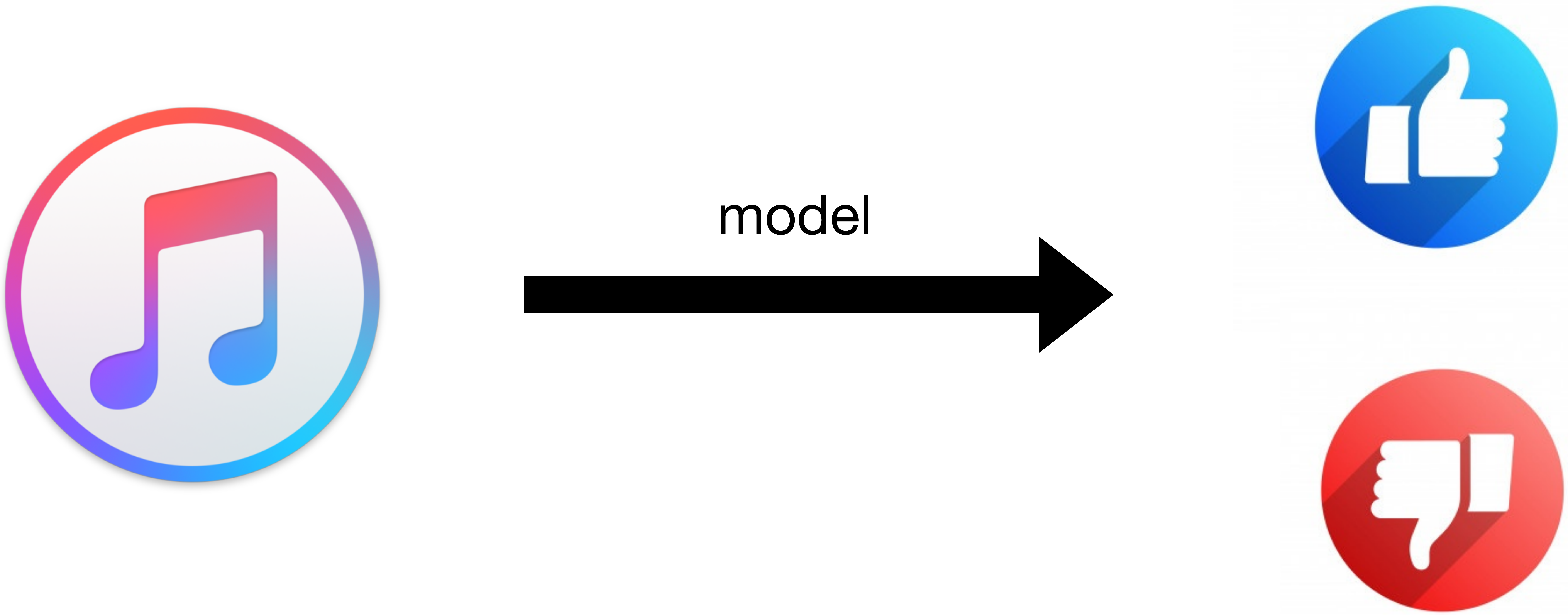
k-nearest neighbors algorithm

From Wikipedia, the free encyclopedia

Not to be confused with k-means clustering.

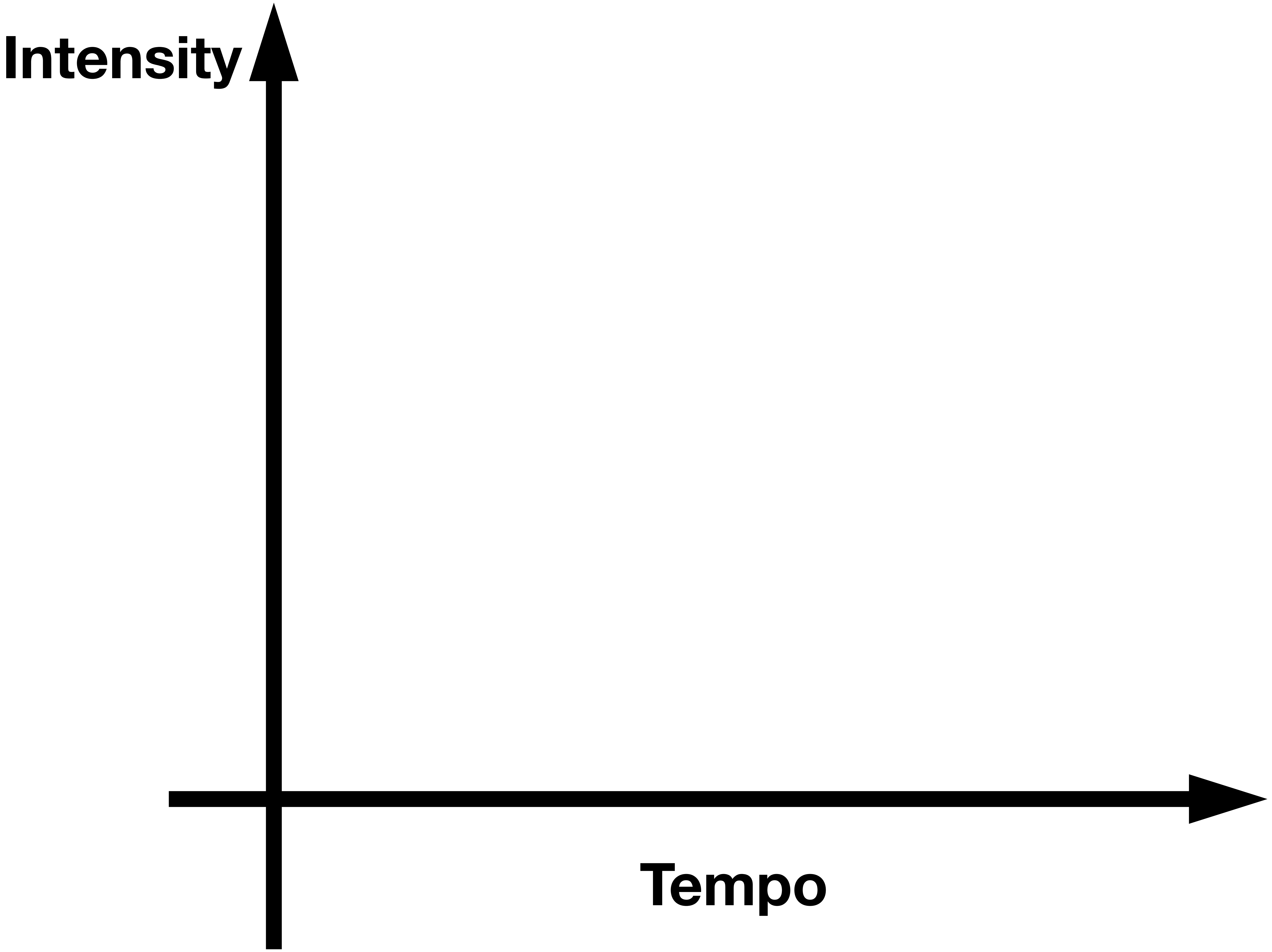
(source: wiki)

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

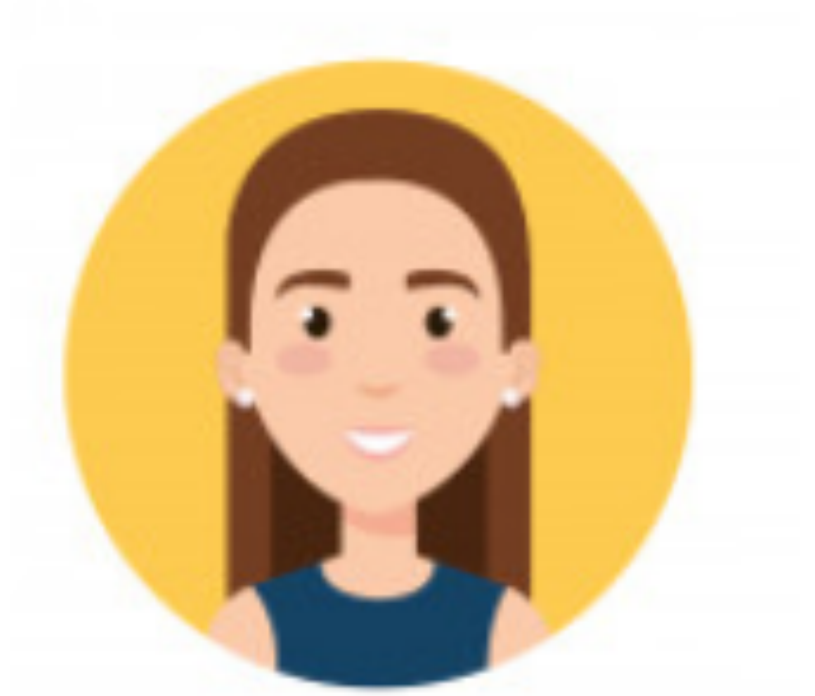


Example 1: Predict whether a user likes a song or not

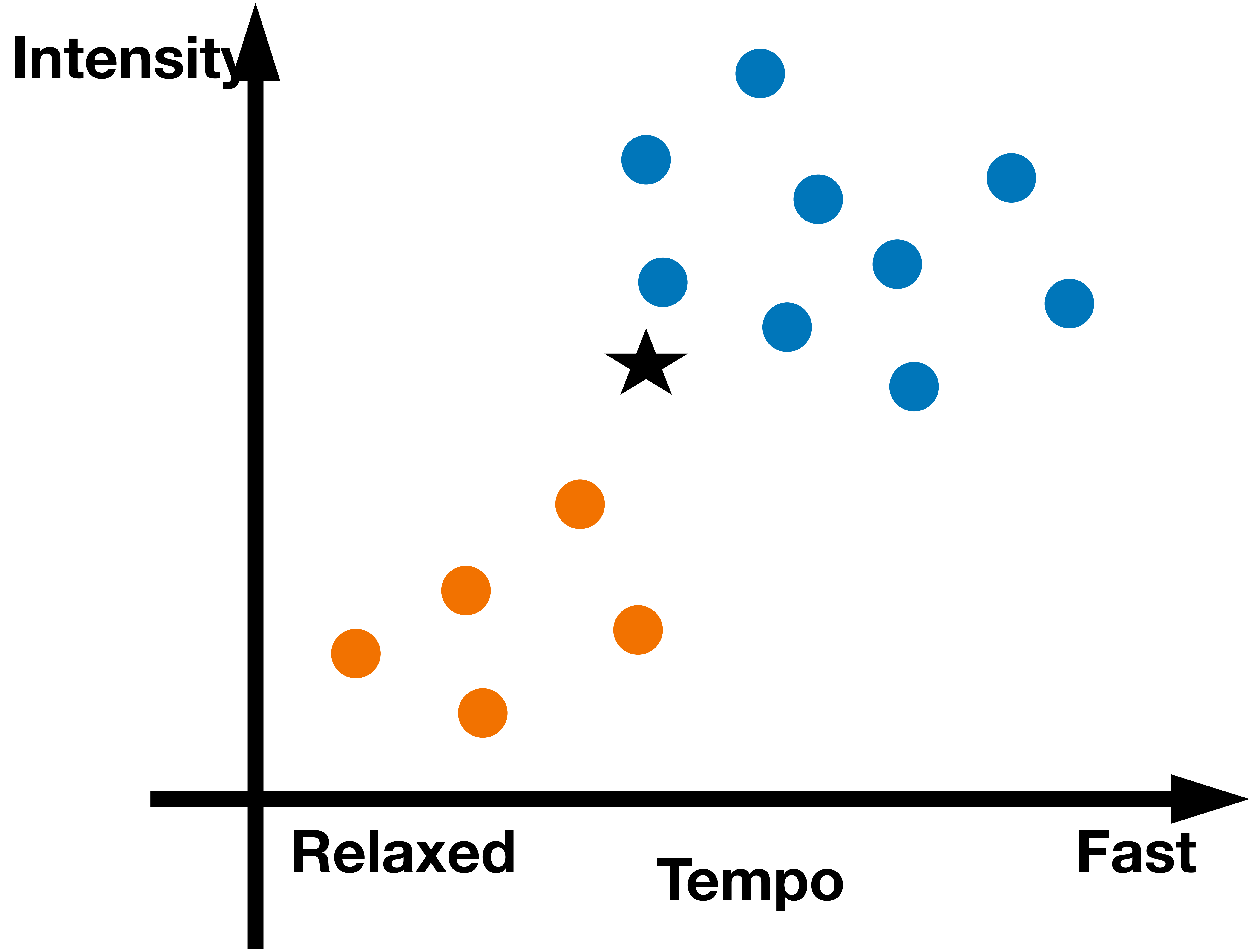
User Sharon



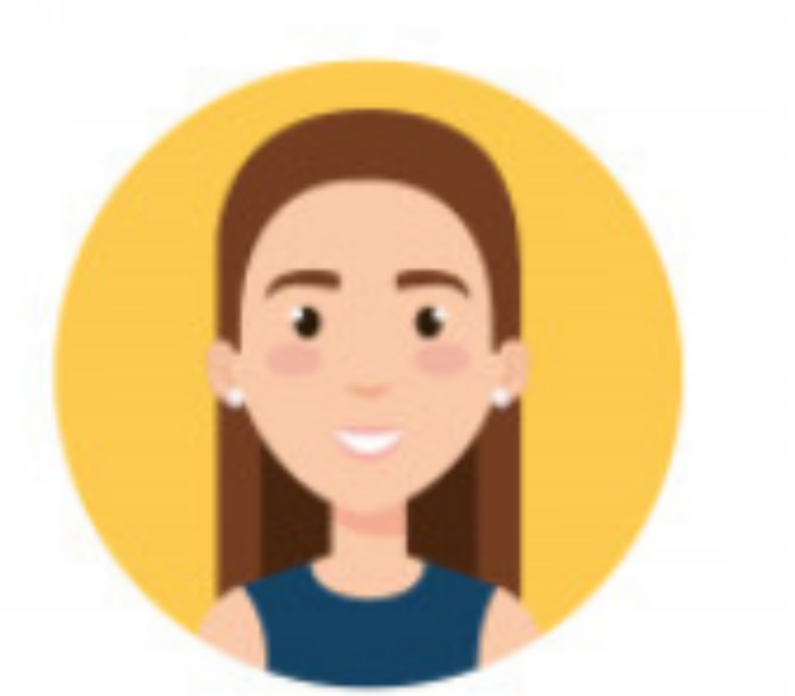
Example 1: Predict whether a user likes a song or not
1-NN



User Sharon



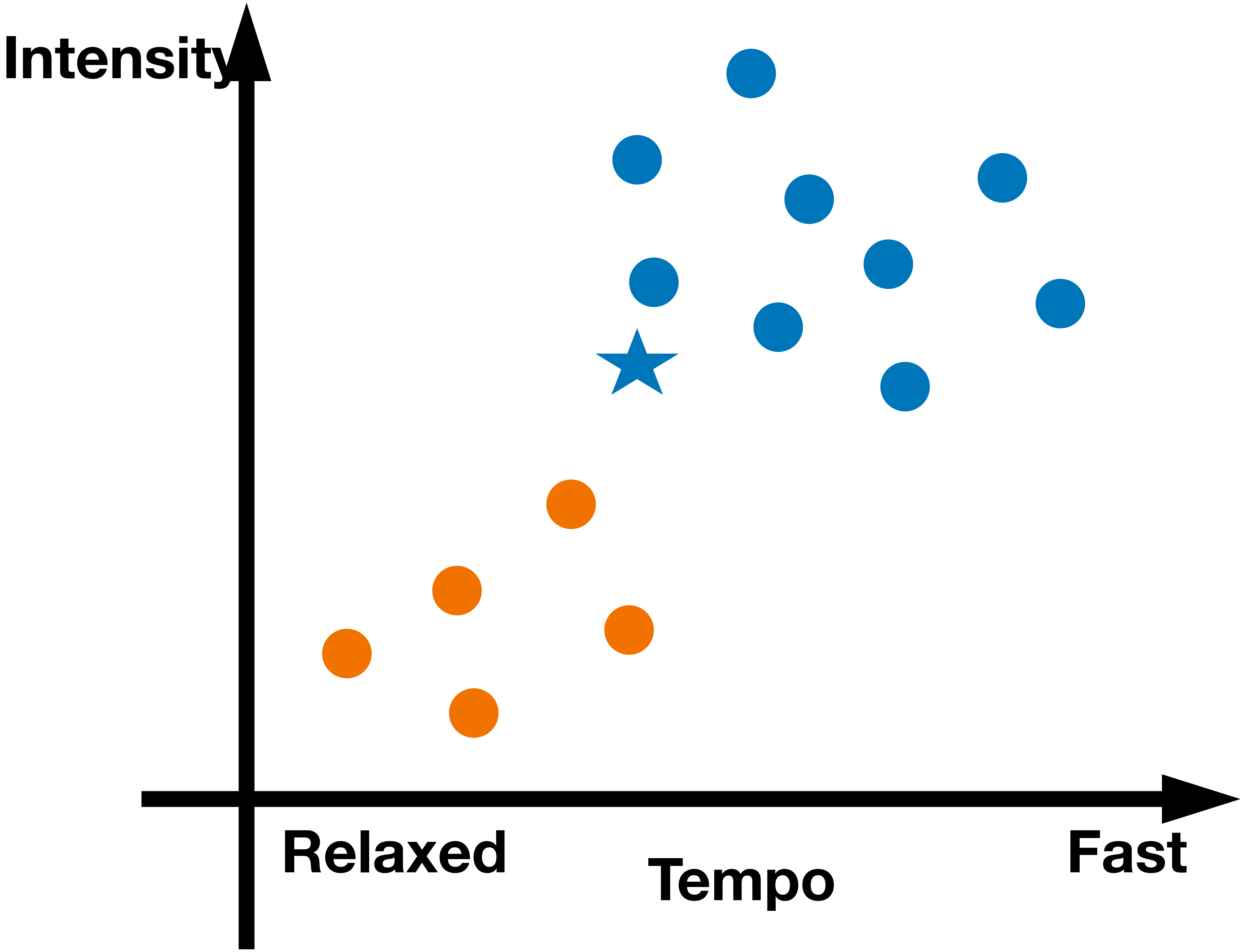
Example 1: Predict whether a user likes a song or not
1-NN



User Sharon

DisLike

Like



K-nearest neighbors for classification

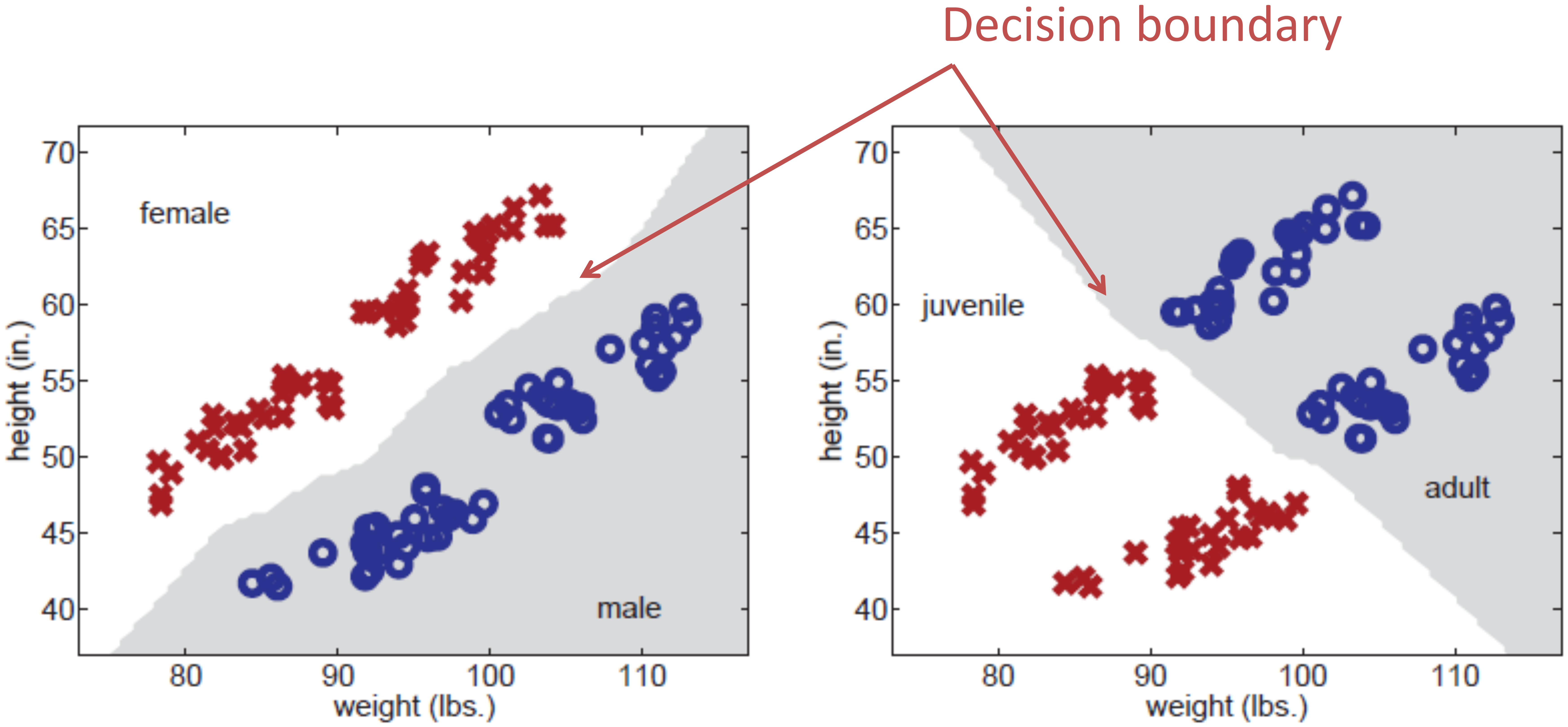
- **Input:** Training data $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$

  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

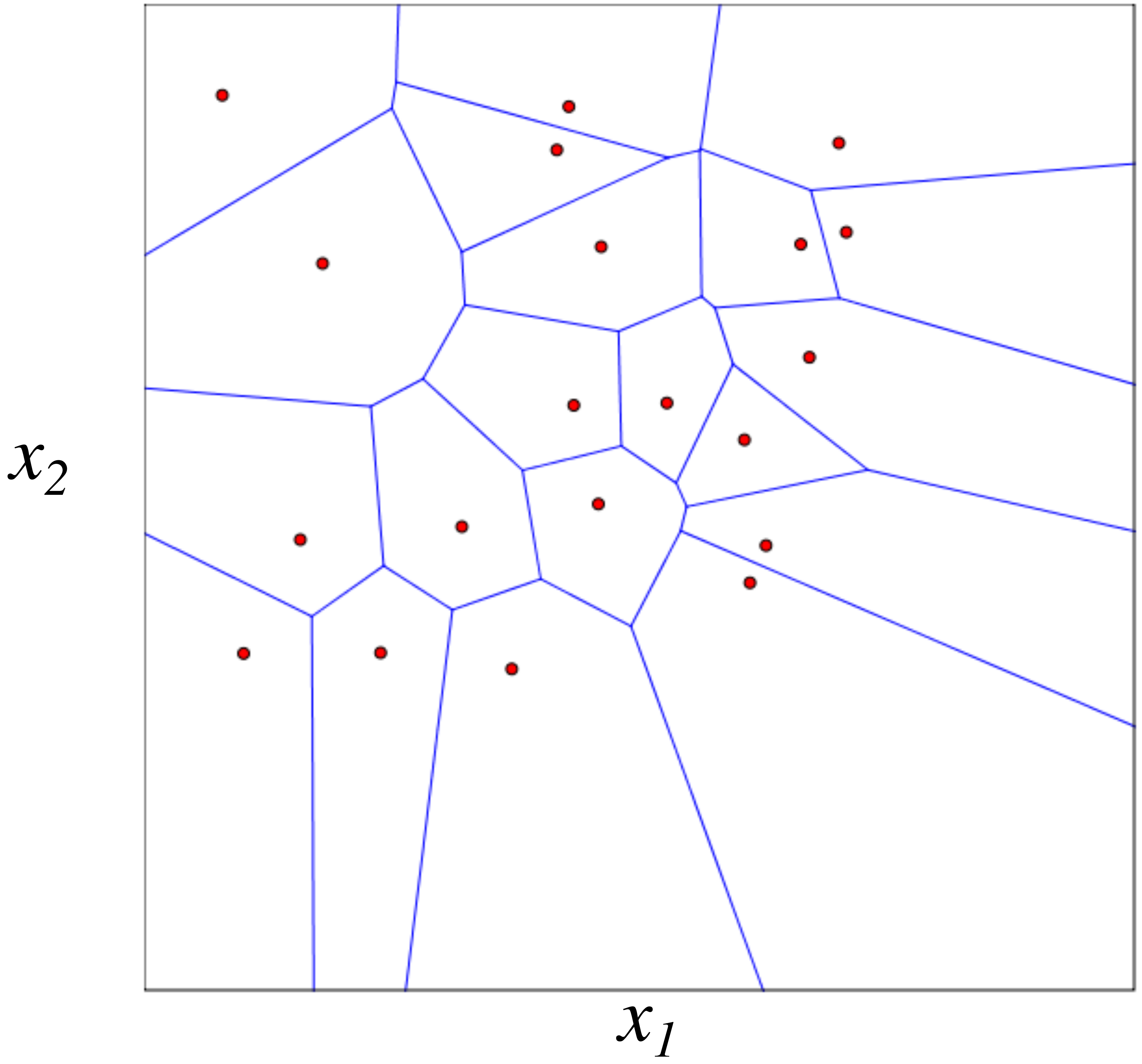


(a) classification by gender

(b) classification by age

The decision regions for 1-NN

Voronoi diagram: each polyhedron indicates the region of feature space that is in the nearest neighborhood of each training instance



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?

suppose all features are discrete

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

suppose all features are continuous

- Euclidean distance: sum of squared differences

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

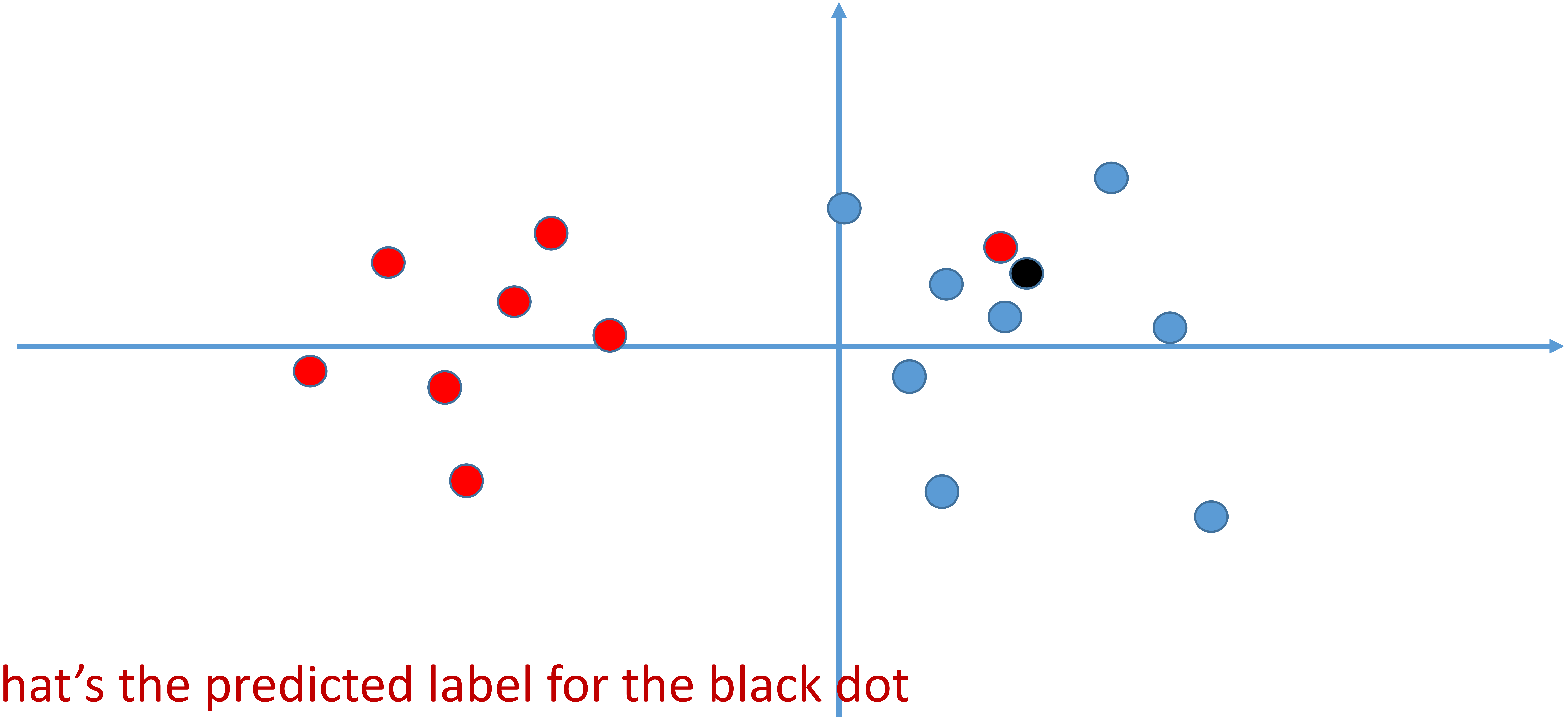
- Manhattan distance:

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

How to pick the number of neighbors

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

Effect of k



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

Part II: Maximum Likelihood Estimation

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)

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)

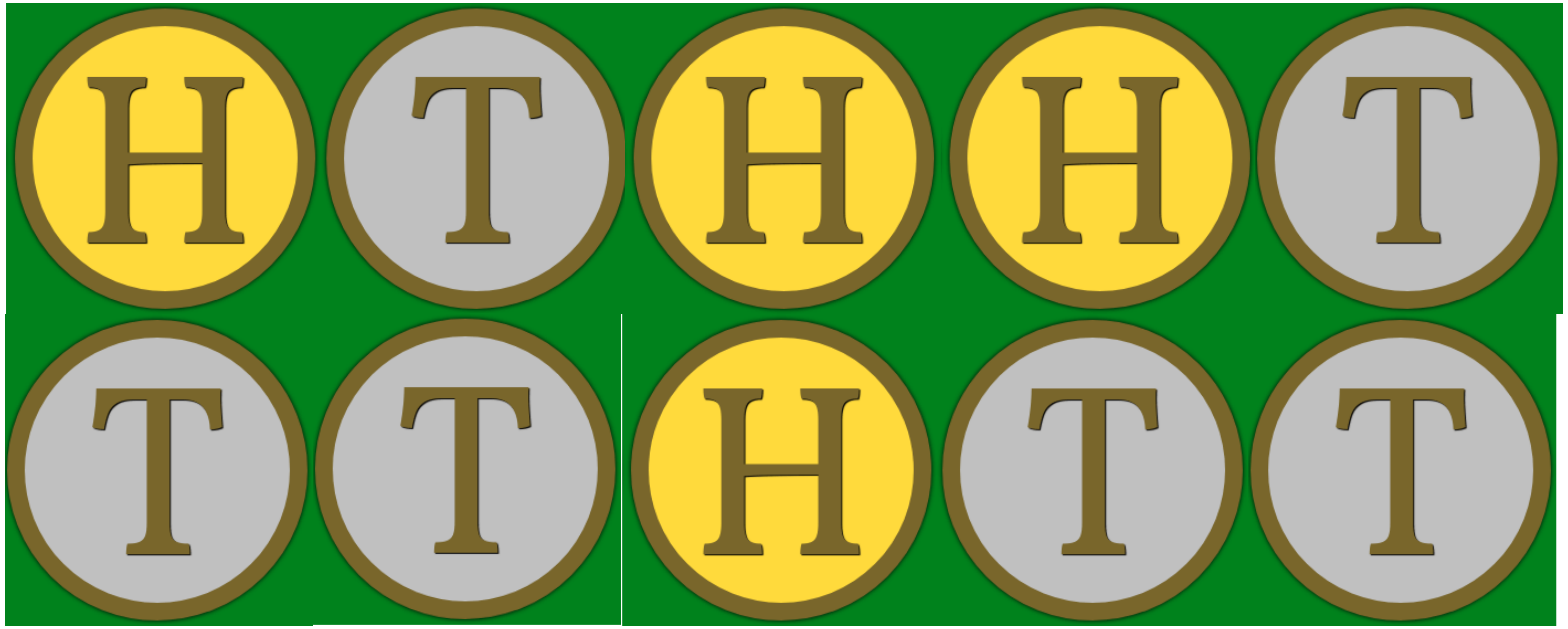
select $\hat{f}$ from a pool of models $\mathcal{F}$
that **minimizes** label disagreement
of the training data

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 $\theta = p(\text{Head})$?



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

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

Under i.i.d assumption

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

How good is θ?

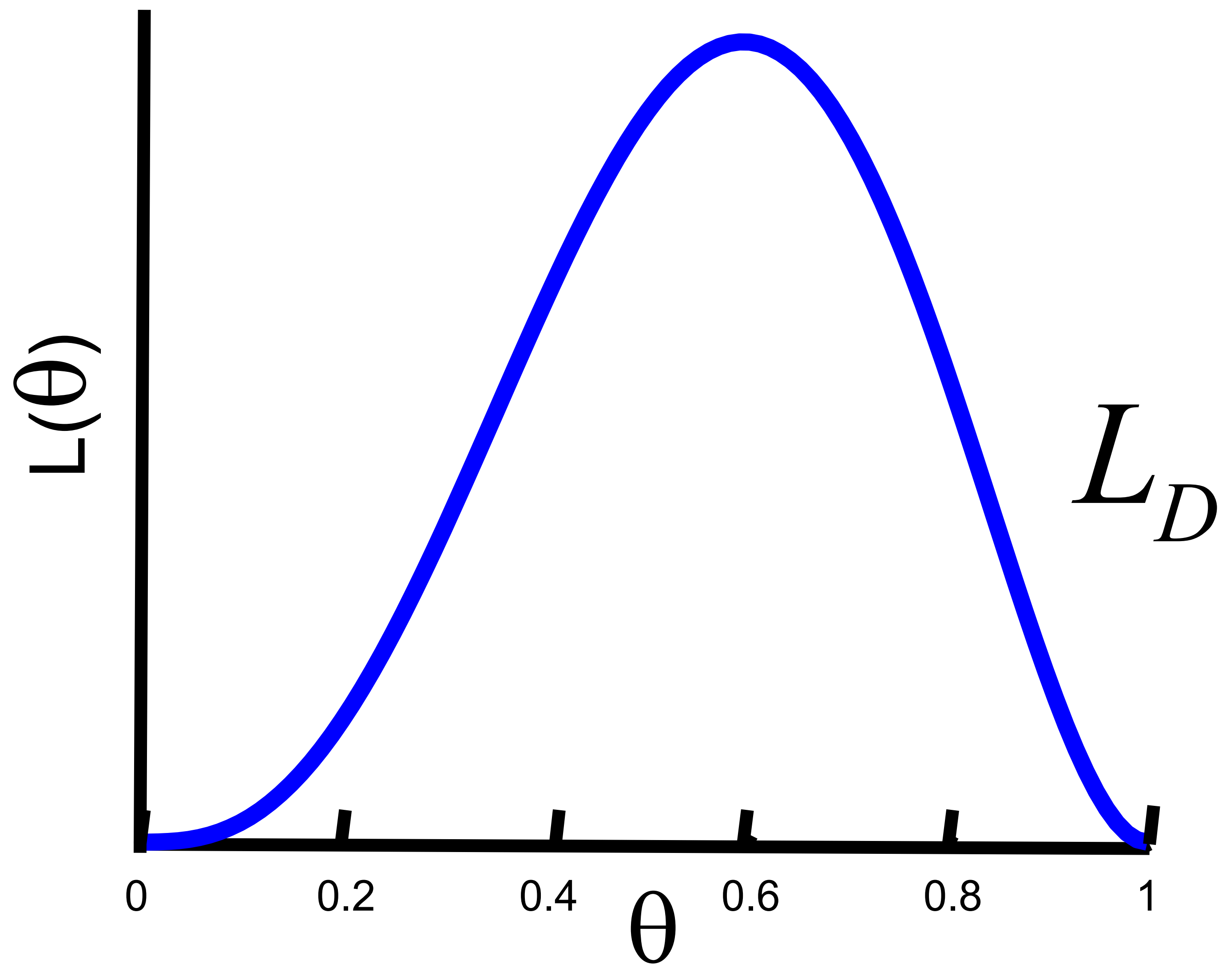
It depends on how likely it is to generate the observed data

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

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

H,T, T, H, H

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



Bernoulli distribution

Log-likelihood function

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

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

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)

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

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

which confirms your intuition!

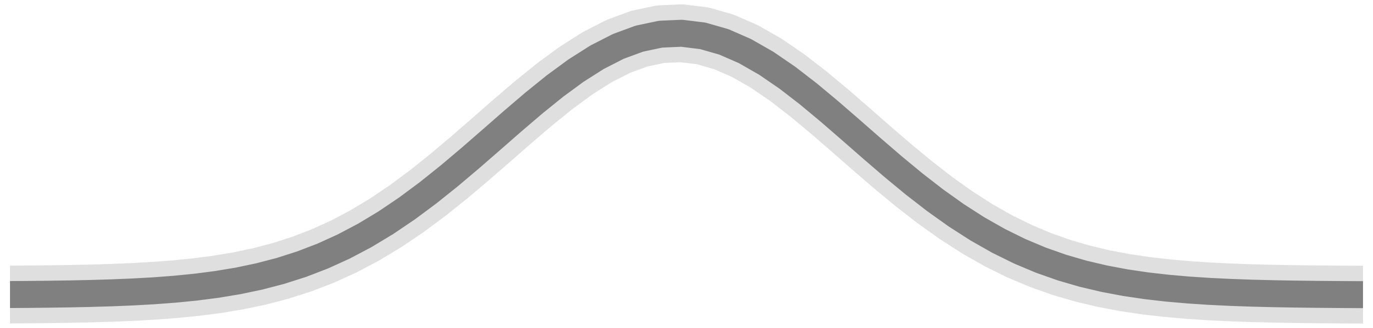
Maximum Likelihood Estimation: Gaussian Model

Fitting a model to heights of females

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

$$\{x_1, x_2, \ldots, x_n\}$$

Model class: Gaussian model



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

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**

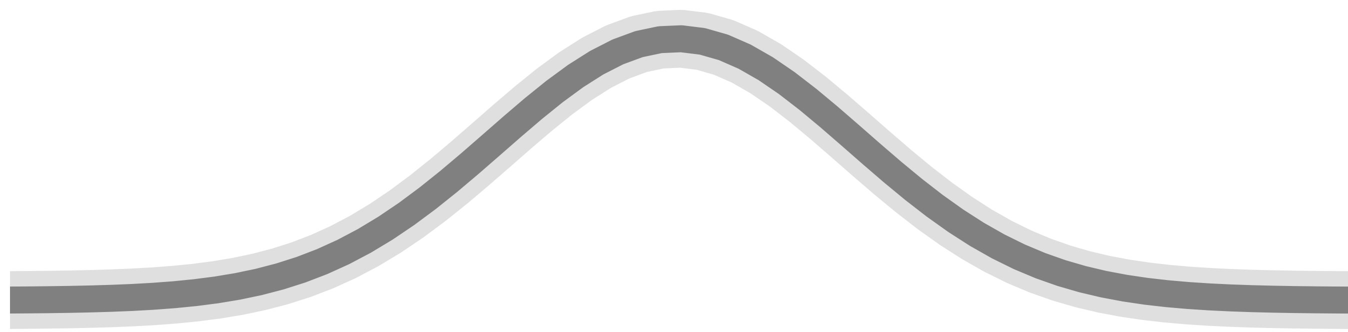
$$\sigma^2 = \mathbf{E}\left[(x-\mu)^2\right] \text{ hence } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

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



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

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

$$\operatorname*{arg\,max}_{\mu,\sigma^2} \prod_{i=1}^{n} p(x_i;\mu,\sigma^2) = \operatorname*{arg\,min}_{\mu,\sigma^2} -\log \prod_{i=1}^{n} p(x_i;\mu,\sigma^2)$$

- **Decompose likelihood**

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

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

Maximum Likelihood

- Estimating the variance

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

Maximum Likelihood

- Estimating the variance

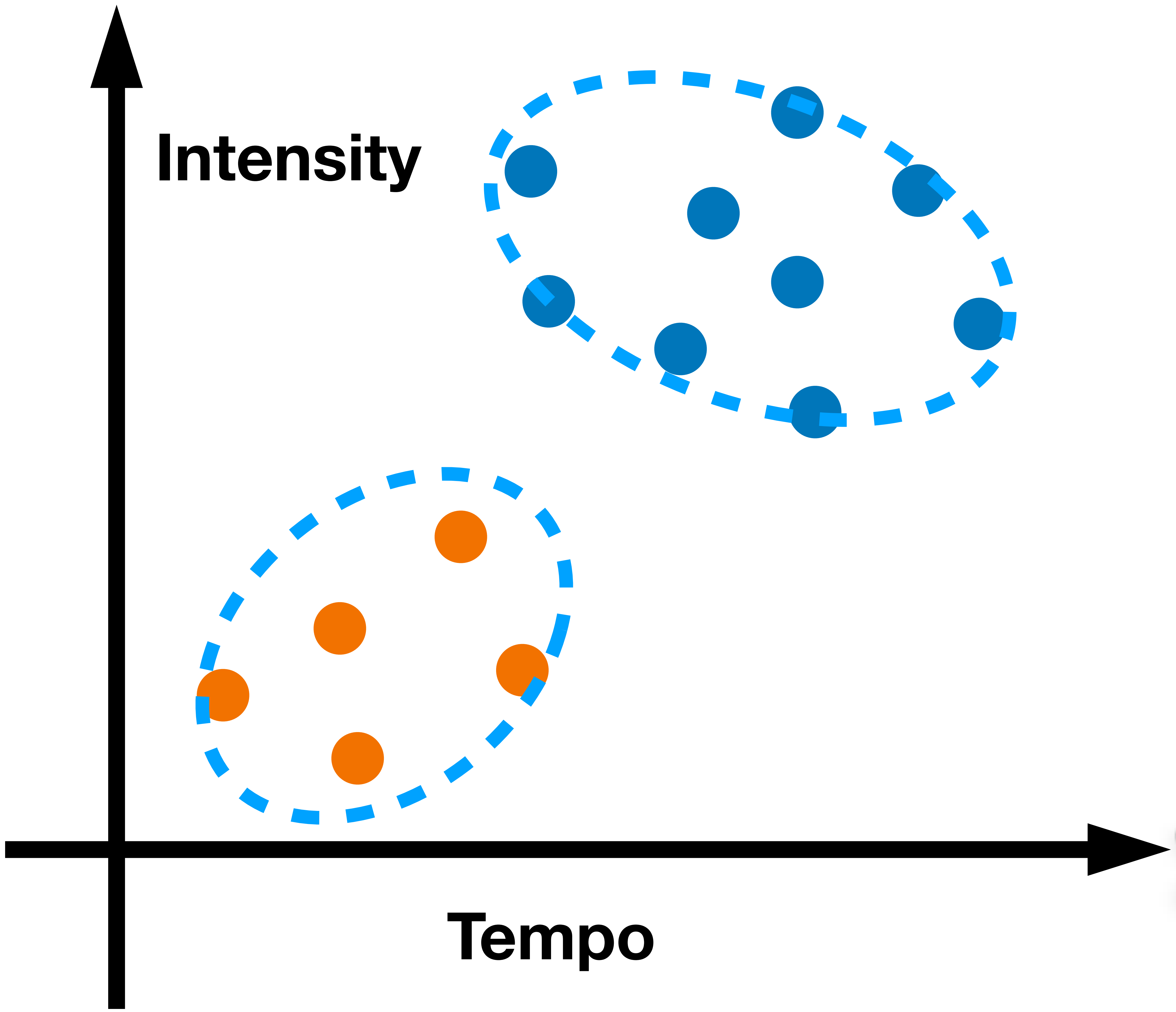
$$\frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

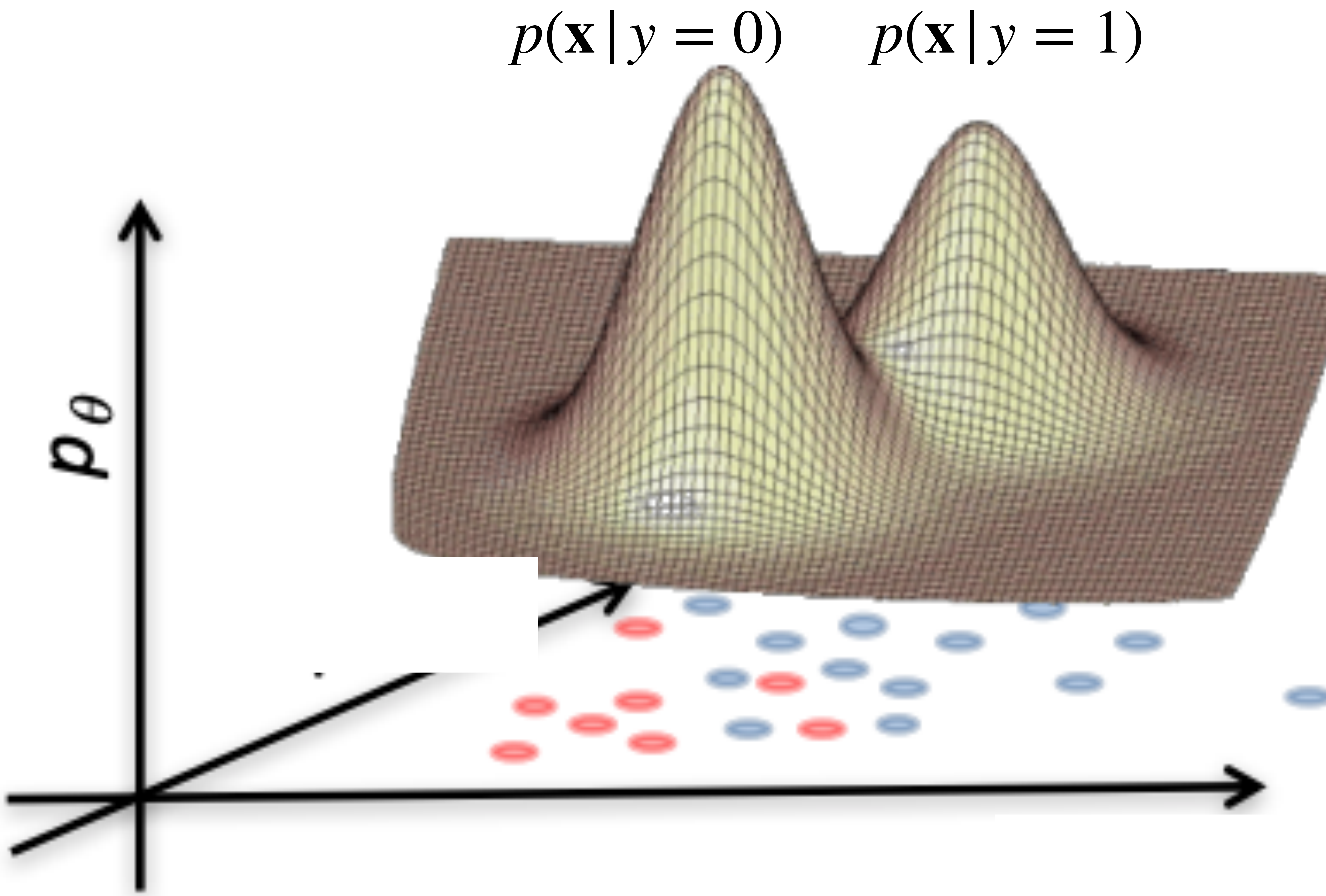
- Take derivatives with respect to it

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

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

Classification via MLE





Classification via MLE

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

(Posterior)

(Prediction)

Classification via MLE

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

(Prediction)

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

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

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

Part II: Naïve Bayes

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}

$$p(\text{Play} \mid ☀) = \frac{p(☀ \mid \text{Play})\, p(\text{Play})}{p(☀)}$$

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(Play = 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!

Bayesian classification

$$\boxed{\hat{y}} = \arg\max \boxed{p(y \mid \mathbf{x})}$$ (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 $\mathbf{x} = \{X_1, \ldots, X_k\}$

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

(Prediction)

(Posterior)

Bayesian classification

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

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

(Prediction)

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

Independent of y

Bayesian classification

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

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

(Prediction)

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

$$= \arg\max_y \underbrace{p(X_1, \ldots, X_k \mid y)}_{\text{Class conditional likelihood}} \; p(y)$$

Class conditional likelihood

Class prior

Naïve Bayes Assumption

Conditional independence of feature attributes

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

Easier to estimate

(using MLE!)

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 Xiaojin (Jerry) Zhu and Yingyu Liang (http://pages.cs.wisc.edu/~jerryzhu/cs540.html), and James McInerney