THE OHIO STATE UNIVERSITY

CSE 5525: Foundations of Speech and Language Processing

Lecture 2: Binary Classification

Huan Sun (CSE@OSU)

Many thanks to Prof. Greg Durrett @ UT Austin for sharing his slides.

Recall: Binary Classification

- Logistic regression: $P(y = 1|x) = \frac{\exp\left(\sum_{i=1}^{n} w_i x_i\right)}{\left(1 + \exp\left(\sum_{i=1}^{n} w_i x_i\right)\right)}$

  Decision rule: $P(y = 1|x) \geq 0.5 \Leftrightarrow w^\top x \geq 0$

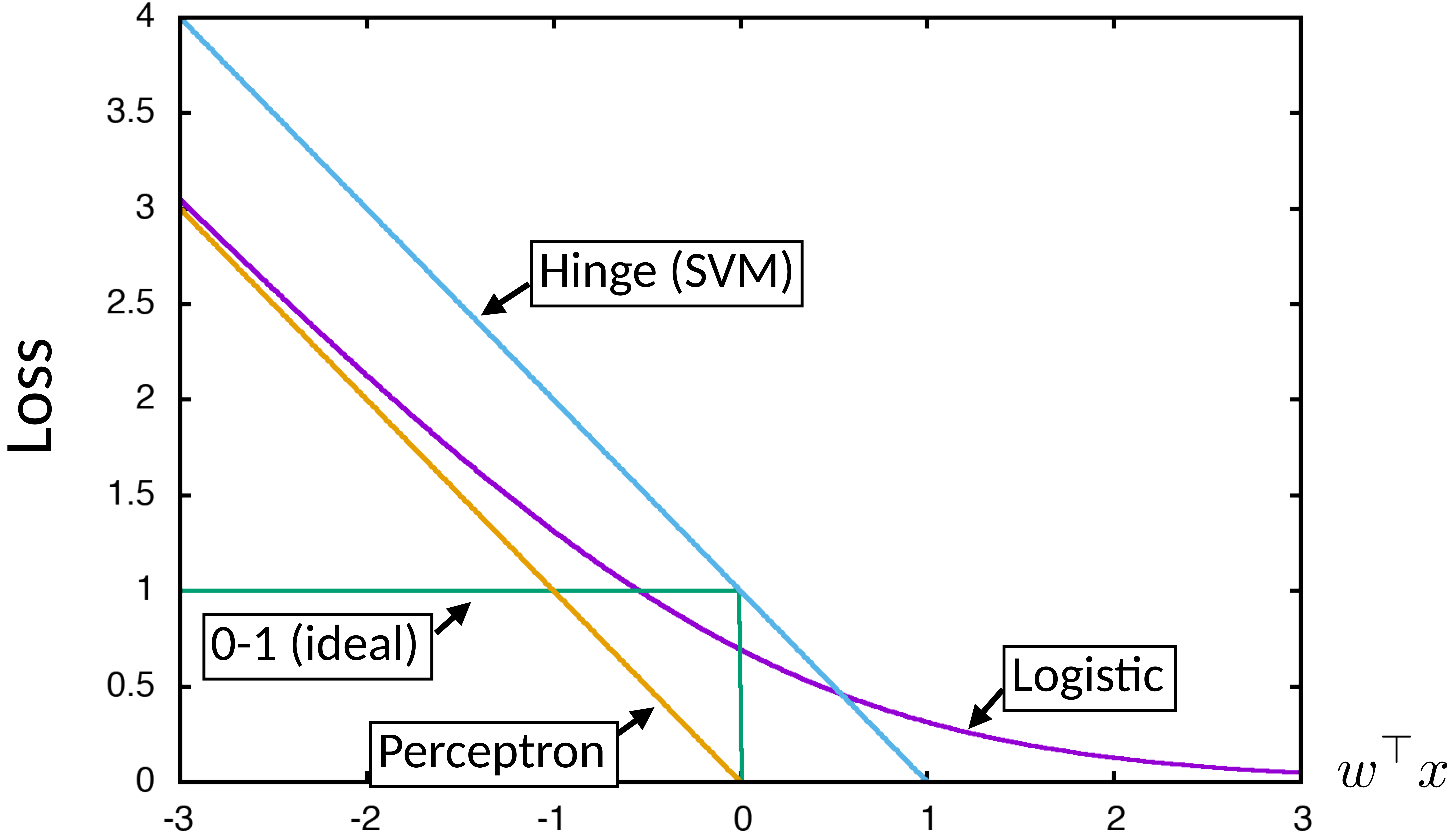
  Gradient (unregularized): $x(y - P(y = 1|x))$

- SVM: quadratic program to minimize weight vector norm w/slack

  Decision rule: $w^\top x \geq 0$

  (Sub)gradient (unregularized): 0 if correct with margin of 1, else $x(2y - 1)$

Loss Functions



This Lecture

- Multiclass fundamentals
- Feature extraction
- Multiclass logistic regression
- Multiclass SVM
- Generative models revisited

"Now! ... *That* should clear up a few things around here!"

Multiclass Fundamentals

Text Classification

A Cancer Conundrum: Too Many Drug Trials, Too Few Patients

Breakthroughs in immunotherapy and a rush to develop profitable new treatments have brought a crush of clinical trials scrambling for patients.

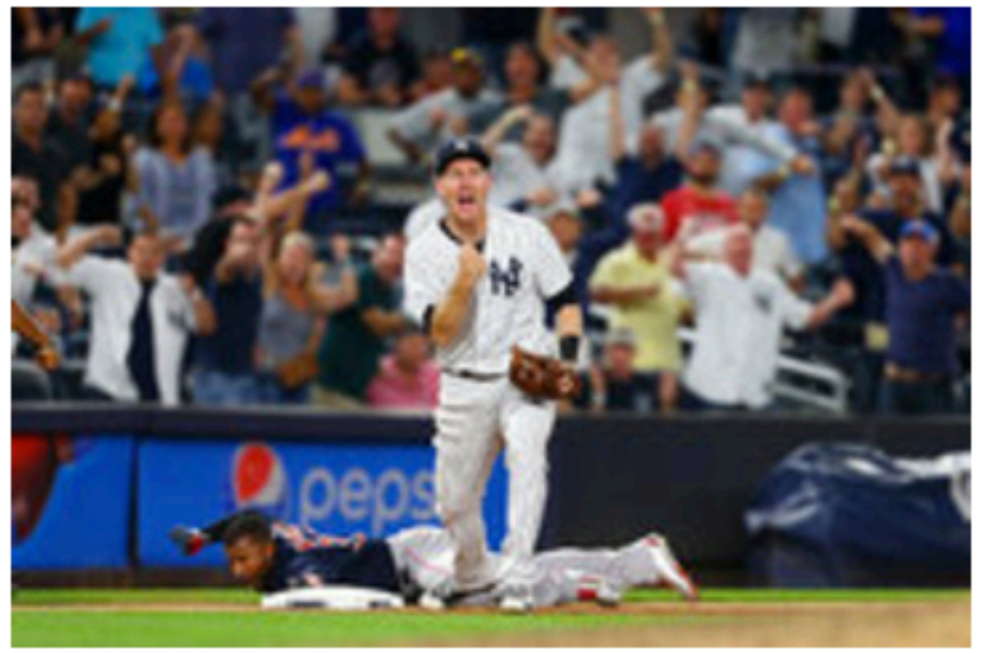
By GINA KOLATA

→ Health

Yankees and Mets Are on Opposite Tracks This Subway Series

As they meet for a four-game series, the Yankees are playing for a postseason spot, and the most the Mets can hope for is to play spoiler.

By FILIP BONDY



→ Sports

~20 classes

Image Classification

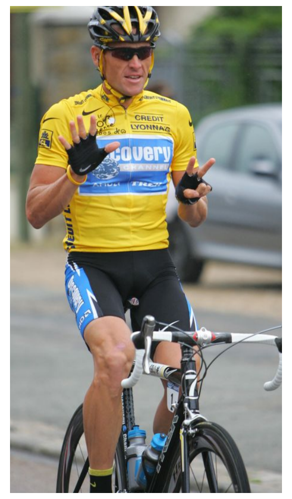
Dog

Car

- Thousands of classes (ImageNet)

Entity Linking

Although he originally won the event, the United States Anti-Doping Agency announced in August 2012 that they had disqualified **Armstrong** from his seven consecutive Tour de France wins from 1999–2005.



Lance Edward Armstrong is an American former professional road cyclist

Armstrong County is a county in Pennsylvania…

?

?

- 4,500,000 classes (all articles in Wikipedia)

Reading Comprehension

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

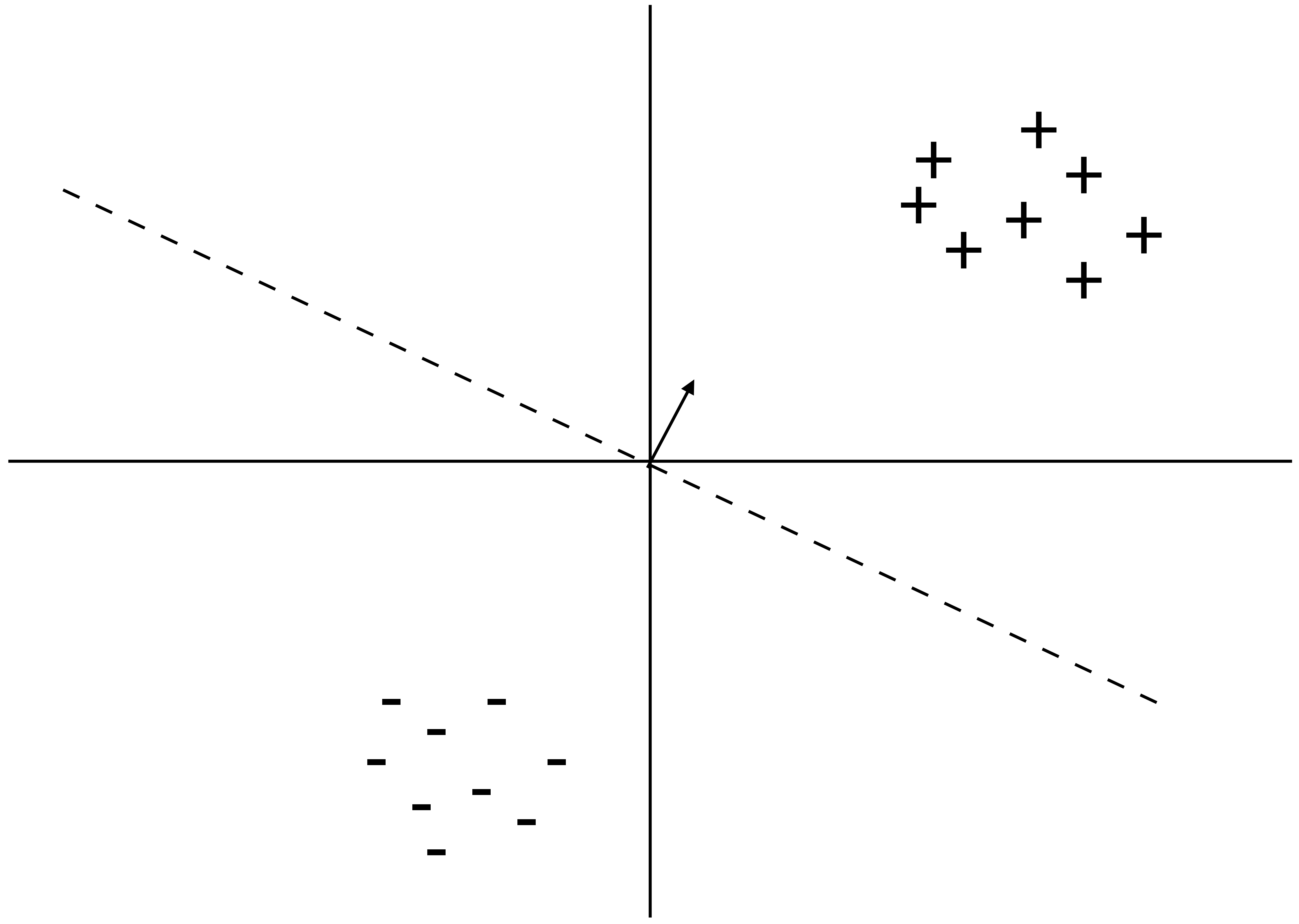
3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room

- Multiple choice questions, 4 classes (but classes change per example)

Richardson (2013)

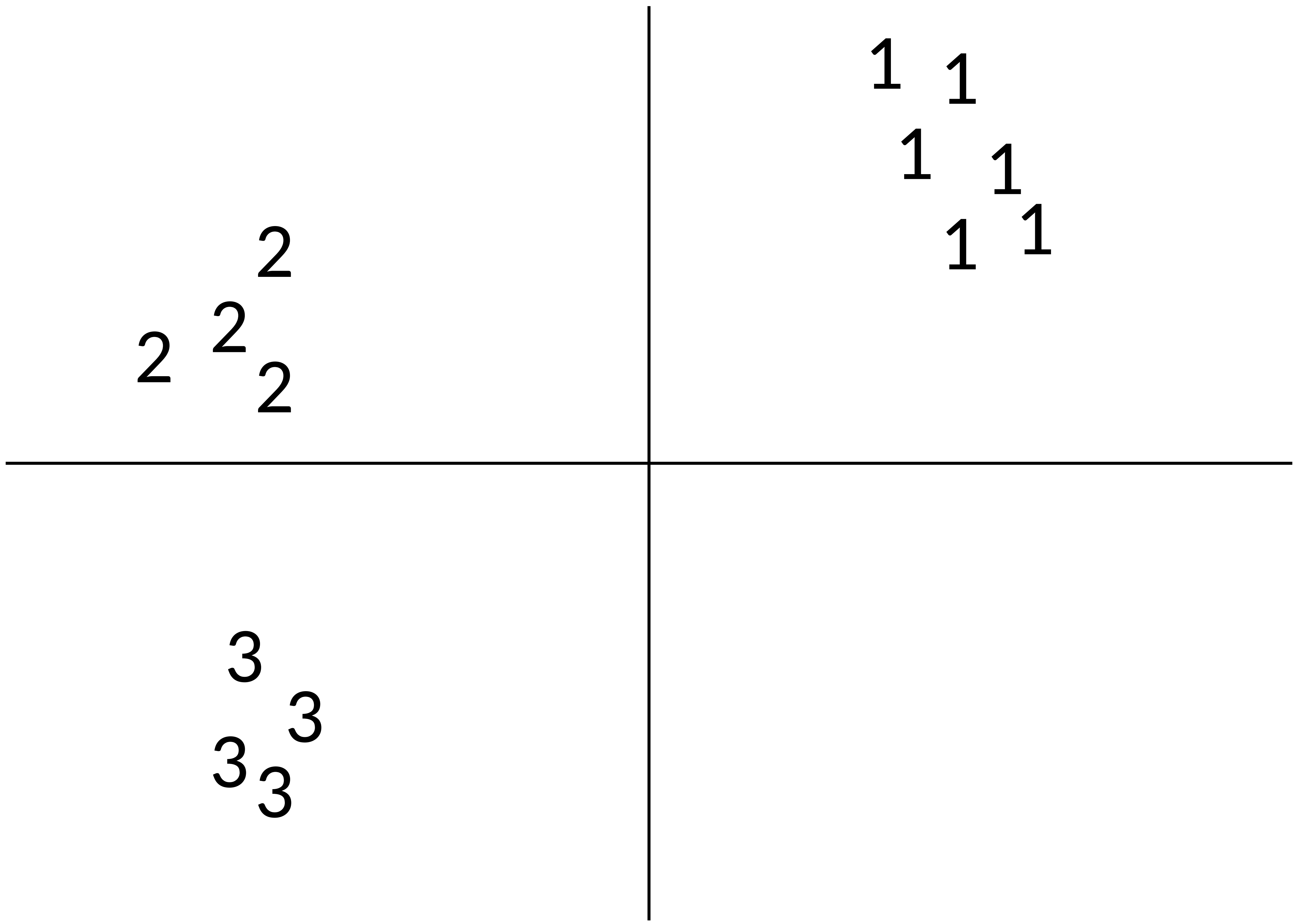
Binary Classification

- Binary classification: one weight vector defines positive and negative classes



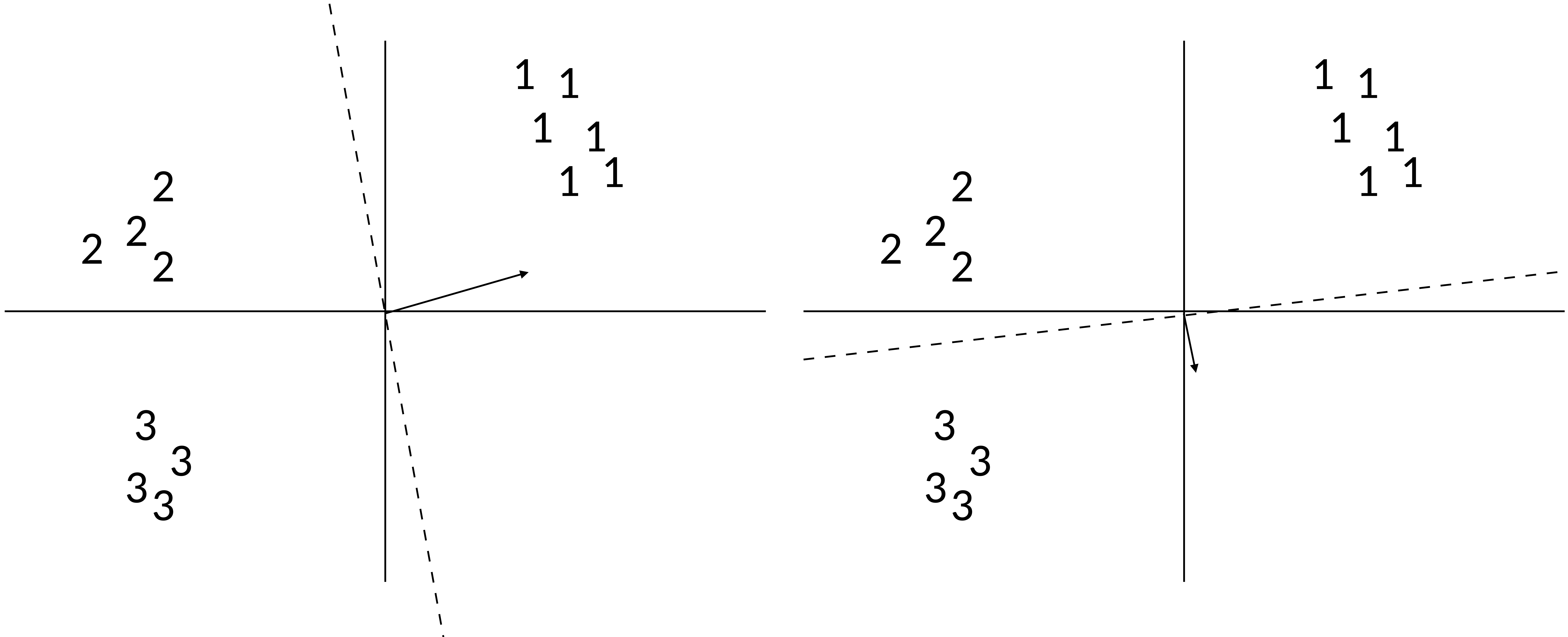
Multiclass Classification

- Can we just use binary classifiers here?



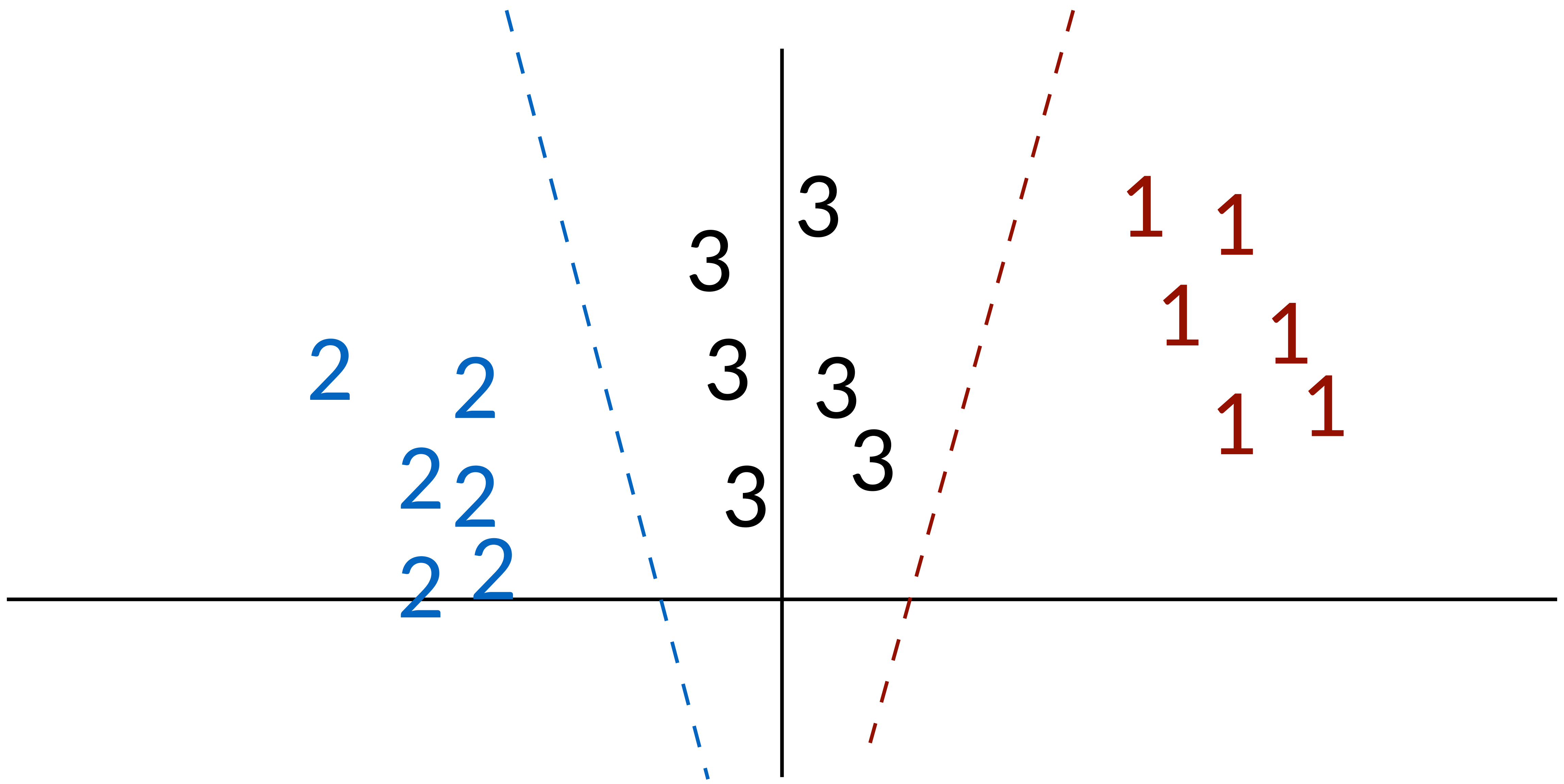
Multiclass Classification

- One-vs-all: train *k* classifiers, one to distinguish each class from all the rest
- How do we reconcile multiple positive predictions? Highest score?



Multiclass Classification

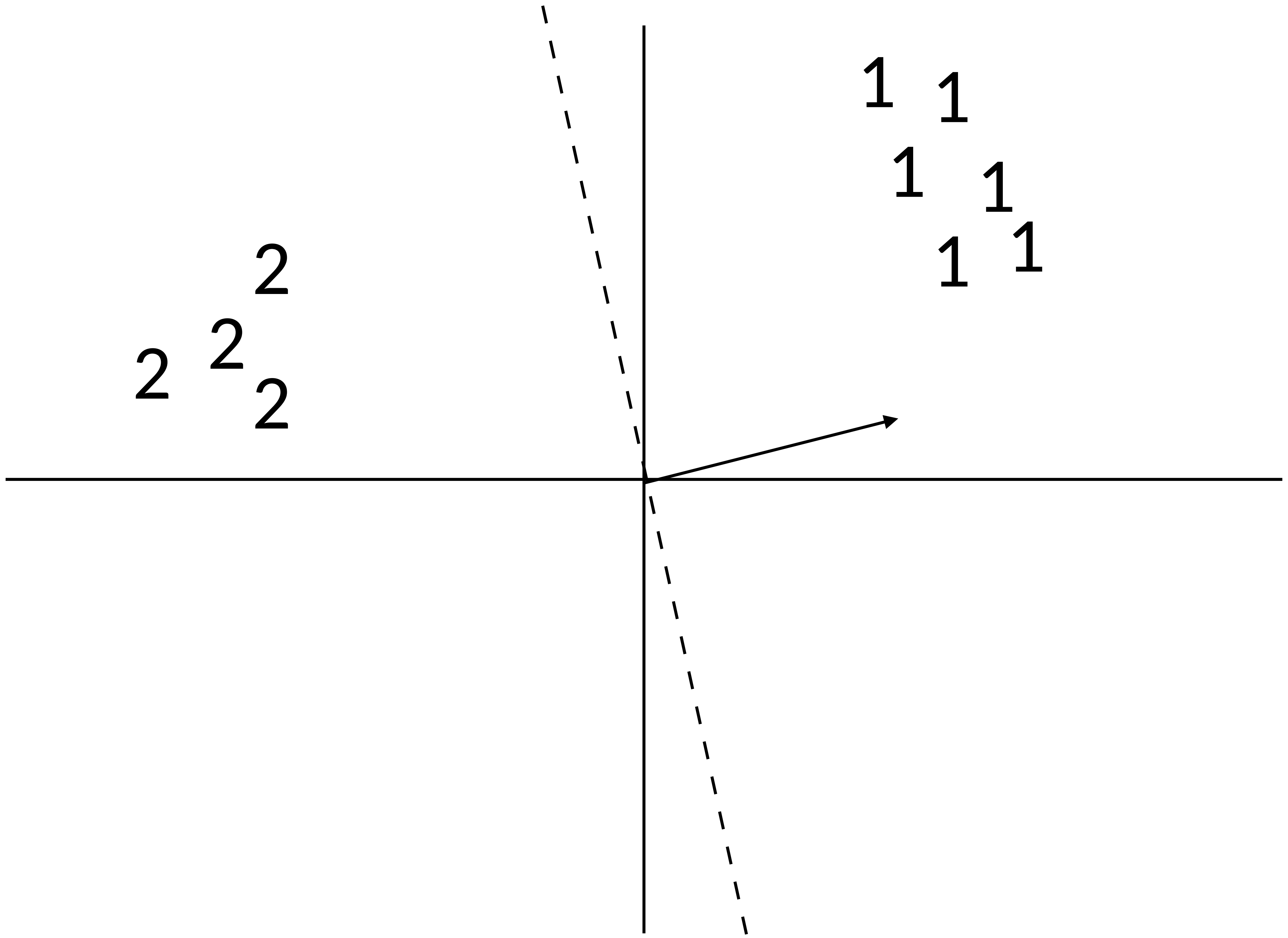
- Not all classes may even be separable using this approach

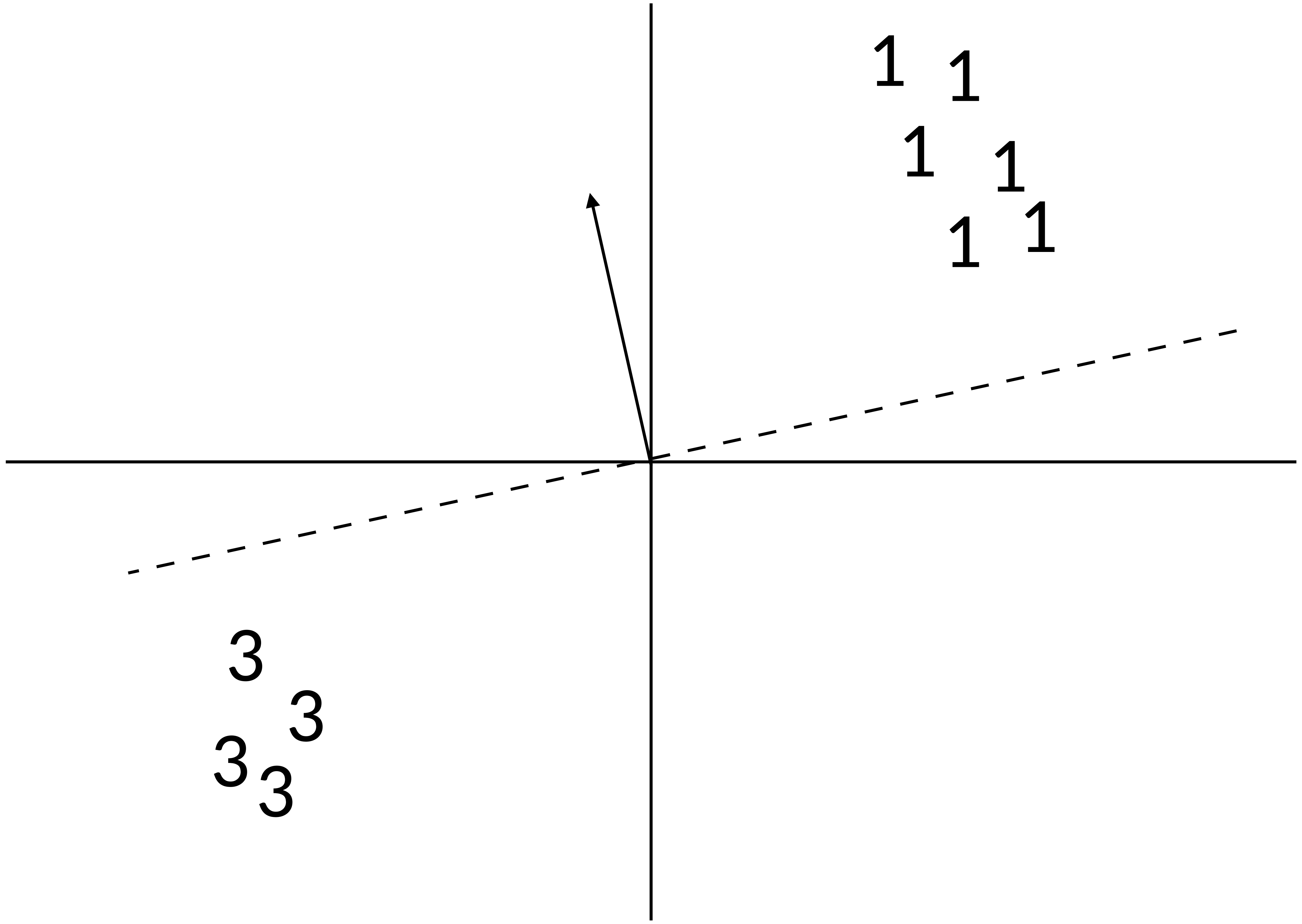


- Can separate 1 from 2+3 and 2 from 1+3 but not 3 from the others (with these features)

Multiclass Classification

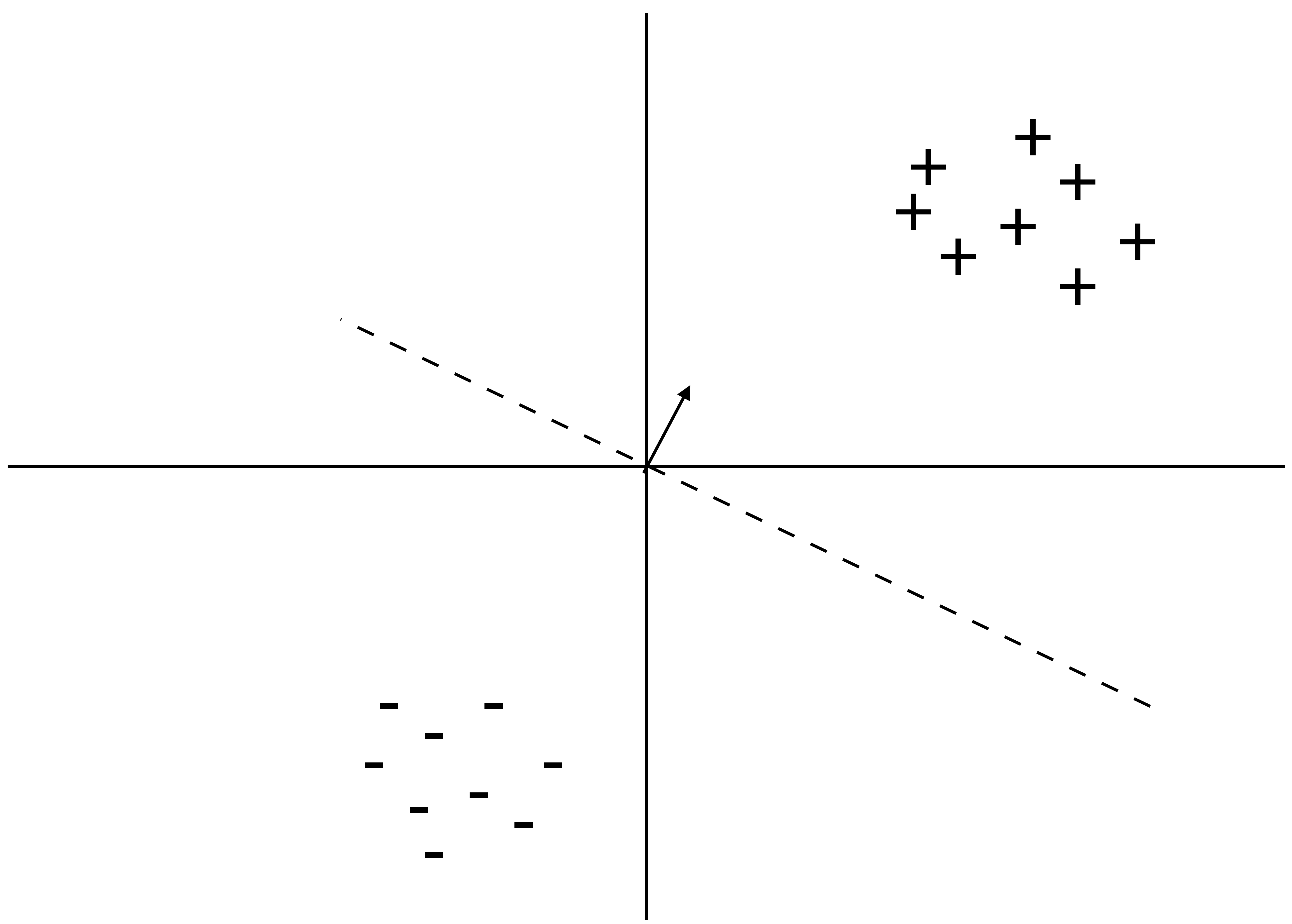
- All-vs-all: train n(n-1)/2 classifiers to differentiate each pair of classes
- Again, how to reconcile?

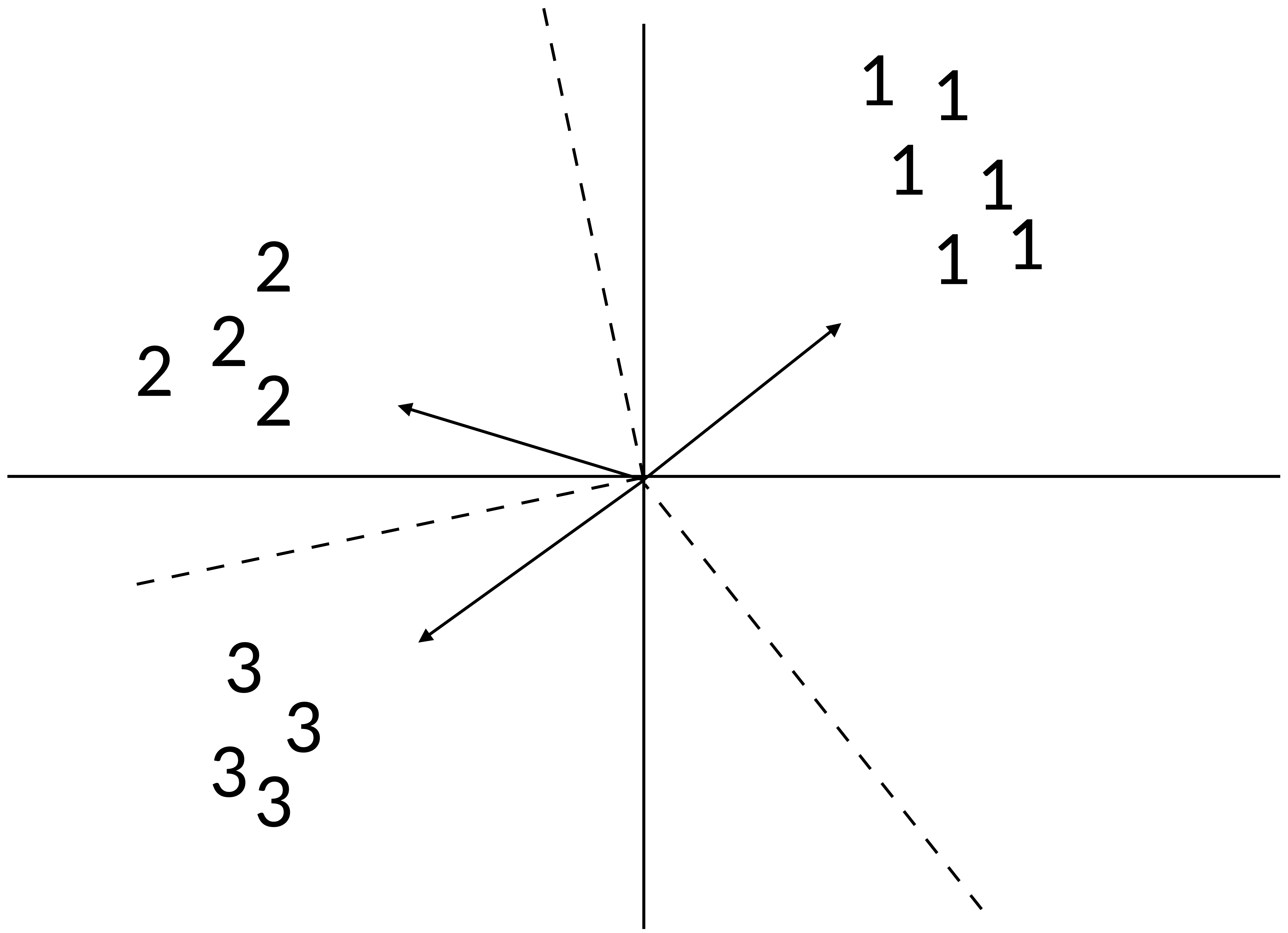




Multiclass Classification

- Binary classification: one weight vector defines both classes
- Multiclass classification: different weights and/or features per class





Multiclass Classification

- Formally: instead of two labels, we have an output space $\mathcal{Y}$ containing a number of possible classes
  - Same machinery that we'll use later for exponentially large output spaces, including sequences and trees

- Decision rule: $\text{argmax}_{y \in \mathcal{Y}} w^\top f(x, y)$

features depend on choice of label now! note: this isn't the gold label

  - Multiple feature vectors, one weight vector

- Can also have one weight vector per class: $\text{argmax}_{y \in \mathcal{Y}} w_y^\top f(x)$

Different Weights vs. Different Features

- Different features: $\mathrm{argmax}_{y \in \mathcal{Y}} w^\top f(x, y)$
  - Suppose $\mathcal{Y}$ is a structured label space (part-of-speech tags for each word in a sentence). *f*(*x*,*y*) extracts features over shared parts of these
- Different weights: $\mathrm{argmax}_{y \in \mathcal{Y}} w_y^\top f(x)$
  - Generalizes to neural networks: *f*(*x*) is the first *n*-1 layers of the network, then you multiply by a final linear layer at the end
- For linear multiclass classification with discrete classes, these are identical

Feature Extraction

Block Feature Vectors

- Decision rule: $\text{argmax}_{y \in \mathcal{Y}} w^\top f(x, y)$

"too many drug trials, too few patients"

Health

Sports

Science

- Base feature function:

$f(x)$ = I[contains *drug*], I[contains *patients*], I[contains *baseball*] = [1, 1, 0]

feature vector blocks for each label

$f(x, y =$ Health$) =$ [1, 1, 0, 0, 0, 0, 0, 0, 0]

I[contains *drug* & label = Health]

$f(x, y =$ Sports$) =$ [0, 0, 0, 1, 1, 0, 0, 0, 0]

- Equivalent to having three weight vectors in this case
- We are NOT looking at the gold label! Instead looking at the candidate label

Making Decisions

too many drug trials, too few patients

Health

Sports

Science

$f(x)$ = I[contains *drug*], I[contains *patients*], I[contains *baseball*]

$f(x, y =$ Health$) =$ [1, 1, 0, 0, 0, 0, 0, 0, 0]

$f(x, y =$ Sports$) =$ [0, 0, 0, 1, 1, 0, 0, 0, 0]

"word drug in Science article" = +1.1

$w =$ [+2.1, +2.3, -5, -2.1, -3.8, +5.2, +1.1, -1.7, -1.3]

$w^\top f(x, y)$ = Health: +4.4 Sports: -5.9 Science: -0.6

argmax

Feature Representation Revisited

this movie was great! would watch again   Positive

[contains *the*] [contains *a*] [contains *was*] [contains *movie*] [contains *film*]

position 0 position 1 position 2 position 3 position 4 ...

- Bag-of-words features are position-insensitive
- What about for tasks like classifying a word as a given part-of-speech?

this movie was great! would watch again

- Want features extracted with respect to this particular position
- [curr word = *was*], [prev word = *movie*], [next word = *great*].
  - How many features?

Multiclass POS tagging

- Classify *blocks* as one of 36 POS tags
- Example *x*: sentence with a word (in this case, *blocks*) highlighted
- Extract features with respect to this word:

the router *blocks* *the packets*

NNS
VBZ
NN
DT
...

$f(x, y{=}\text{VBZ})$ = I[curr_word=*blocks* & tag = VBZ],
I[prev_word=*router* & tag = VBZ]
I[next_word=*the* & tag = VBZ]
I[curr_suffix=*s* & tag = VBZ]

not saying that *the* is tagged as VBZ! saying that *the* follows the VBZ word

- Next two lectures: sequence labeling!

Multiclass Logistic Regression

Multiclass Logistic Regression

$$P_w(y|x) = \frac{\exp\left(w^\top f(x,y)\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(w^\top f(x,y')\right)}$$

sum over output space to normalize

- Compare to binary:

$$P(y=1|x) = \frac{\exp(w^\top f(x))}{1+\exp(w^\top f(x))}$$

negative class implicitly had w^T*f*(*x*, *y*=0) = the zero vector

- exp/sum(exp): also called *softmax*
- Training: maximize $\mathcal{L}(x,y) = \sum_{j=1}^{n} \log P(y_j^*|x_j)$

$$= \sum_{j=1}^{n} \left(w^\top f(x_j, y_j^*) - \log \sum_y \exp(w^\top f(x_j, y)) \right)$$

Training

- Multiclass logistic regression $P_w(y|x) = \frac{\exp\left(w^\top f(x,y)\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(w^\top f(x,y')\right)}$

- Likelihood $\mathcal{L}(x_j, y_j^*) = w^\top f(x_j, y_j^*) - \log \sum_y \exp(w^\top f(x_j, y))$

$$\frac{\partial}{\partial w_i}\mathcal{L}(x_j, y_j^*) = f_i(x_j, y_j^*) - \frac{\sum_y f_i(x_j, y)\exp(w^\top f(x_j, y))}{\sum_y \exp(w^\top f(x_j, y))}$$

$$\frac{\partial}{\partial w_i}\mathcal{L}(x_j, y_j^*) = f_i(x_j, y_j^*) - \sum_y f_i(x_j, y) P_w(y|x_j)$$

$$\frac{\partial}{\partial w_i}\mathcal{L}(x_j, y_j^*) = f_i(x_j, y_j^*) - \mathbb{E}_y[f_i(x_j, y)]$$

gold feature value

model's expectation of feature value

Training

$$\frac{\partial}{\partial w_i}\mathcal{L}(x_j, y_j^*) = f_i(x_j, y_j^*) - \sum_y f_i(x_j, y) P_w(y|x_j)$$

too many drug trials, too few patients

y^* = Health

$f(x, y =$ Health$) = [1, 1, 0, 0, 0, 0, 0, 0, 0]$

$f(x, y =$ Sports$) = [0, 0, 0, 1, 1, 0, 0, 0, 0]$

$P_w(y|x) = [0.2, 0.5, 0.3]$
(made up values)

gradient:

[1, 1, 0, 0, 0, 0, 0, 0, 0] — 0.2 [1, 1, 0, 0, 0, 0, 0, 0, 0] — 0.5 [0, 0, 0, 1, 1, 0, 0, 0, 0]
— 0.3 [0, 0, 0, 0, 0, 0, 1, 1, 0]

= [0.8, 0.8, 0, -0.5, -0.5, 0, -0.3, -0.3, 0]

Logistic Regression: Summary

- Model: $P_w(y|x) = \dfrac{\exp\left(w^\top f(x,y)\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(w^\top f(x,y')\right)}$

- Inference: $\text{argmax}_y P_w(y|x)$

- Learning: gradient ascent on the discriminative log-likelihood

$$f(x,y^*) - \mathbb{E}_y[f(x,y)] = f(x,y^*) - \sum_y [P_w(y|x) f(x,y)]$$

"towards gold feature value, away from expectation of feature value"

Multiclass SVM

Loss Functions

- Are all decisions equally costly?

too many drug trials, too few patients → Health, Sports, Science

Predicted Sports : bad error

Predicted Science: not so bad

- We can define a loss function $\ell(y, y^*)$

$$\ell(\text{Sports}, \text{Health}) = 3$$

$$\ell(\text{Science}, \text{Health}) = 1$$

Multiclass SVM

Minimize $\lambda \|w\|_2^2 + \sum_{j=1}^{m} \xi_j$

slack variables > 0 iff example is support vector

s.t. $\forall j \quad \xi_j \geq 0$

$\cancel{\forall j \quad (2y_j - 1)(w^\top x_j) \geq 1 - \xi_j}$

$\forall j \forall y \in \mathcal{Y} \quad w^\top f(x_j, y_j^*) \geq w^\top f(x_j, y) + \ell(y, y_j^*) - \xi_j$

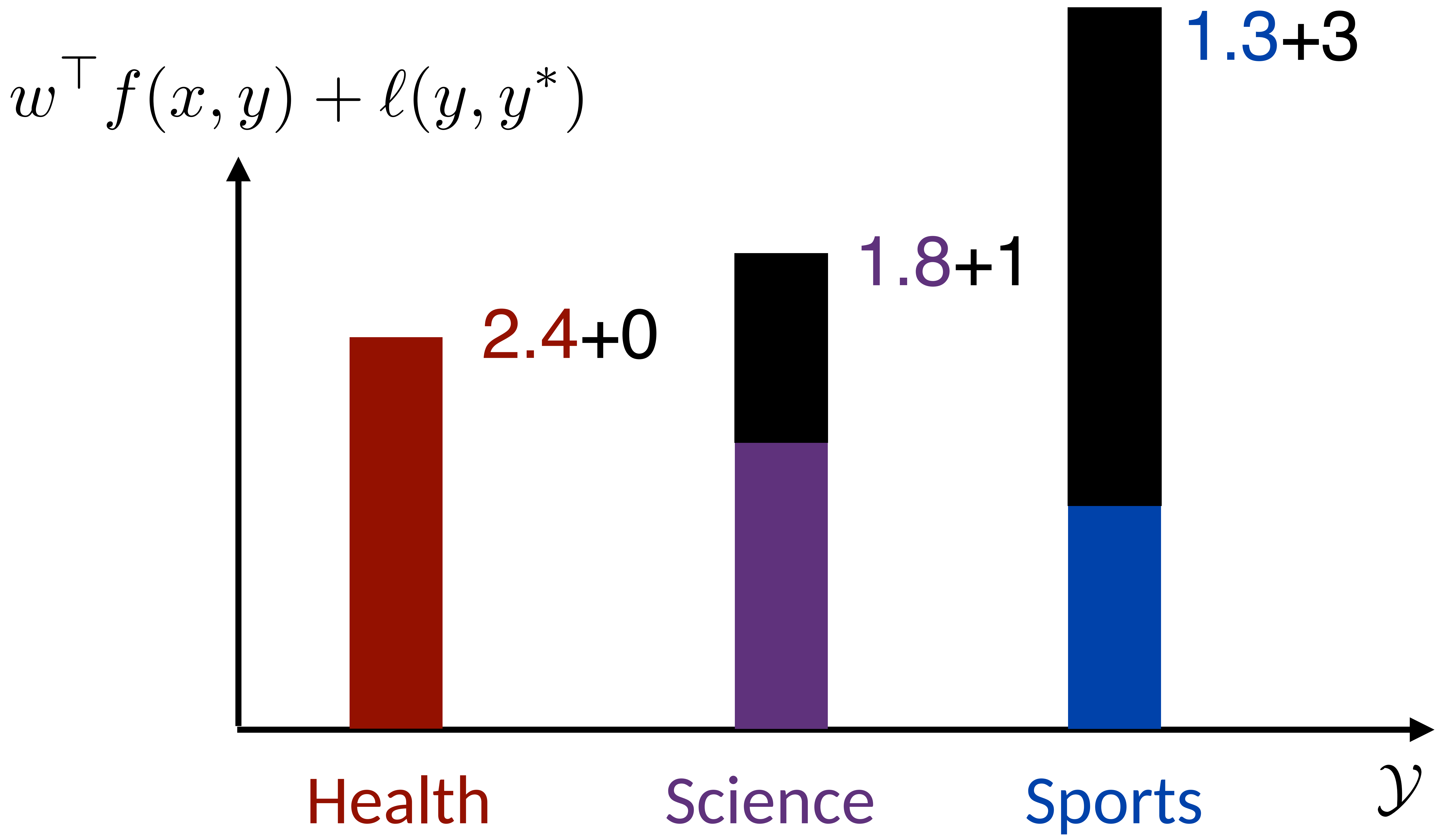
Correct prediction now has to beat every other class

Score comparison is more explicit now

The 1 that was here is replaced by a loss function

Multiclass SVM

$$\forall j \forall y \in \mathcal{Y} \quad w^\top f(x_j, y_j^*) \geq w^\top f(x_j, y) + \ell(y, y_j^*) - \xi_j$$



- Does gold beat every label + loss? No!
- Most violated constraint is Sports; what is ξ_j?
- ξ_j = 4.3 - 2.4 = 1.9
- Perceptron would make no update here

Revisiting Generative vs. Discriminative Models

Learning in Probabilistic Models

- So far we have talked about discriminative classifiers (e.g., logistic regression which models $P(y|x)$)
- Cannot analytically compute optimal weights for such models, need to use gradient descent
- What about generative models?

Naive Bayes

- Data point $x = (x_1, ..., x_n)$, label $y \in \{0, 1\}$
- Formulate a probabilistic model that places a distribution $P(x, y)$
- Compute $P(y|x)$, predict $\text{argmax}_y P(y|x)$ to classify

$$P(y|x) = \frac{P(y)P(x|y)}{P(x)}$$

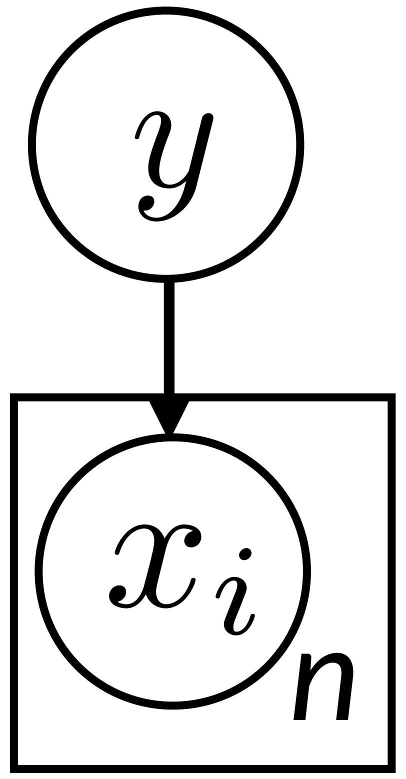
Bayes' Rule

constant: irrelevant
for finding the max

$$\propto P(y)P(x|y)$$

"Naive" assumption:

$$= P(y) \prod_{i=1}^{n} P(x_i|y)$$



Maximum Likelihood Estimation

- Data points (x_j, y_j) provided (*j* indexes over examples)
- Find values of $P(y),\ \ P(x_i|y)$ that maximize data likelihood (generative):

$$\underbrace{\prod_{j=1}^{m}}_{\text{data points }(j)} P(y_j, x_j) = \underbrace{\prod_{j=1}^{m}}_{\text{data points }(j)} P(y_j) \left[\underbrace{\prod_{i=1}^{n}}_{\text{features }(i)} P(\underbrace{x_{ji}}_{i\text{th feature of }j\text{th example}}|y_j) \right]$$

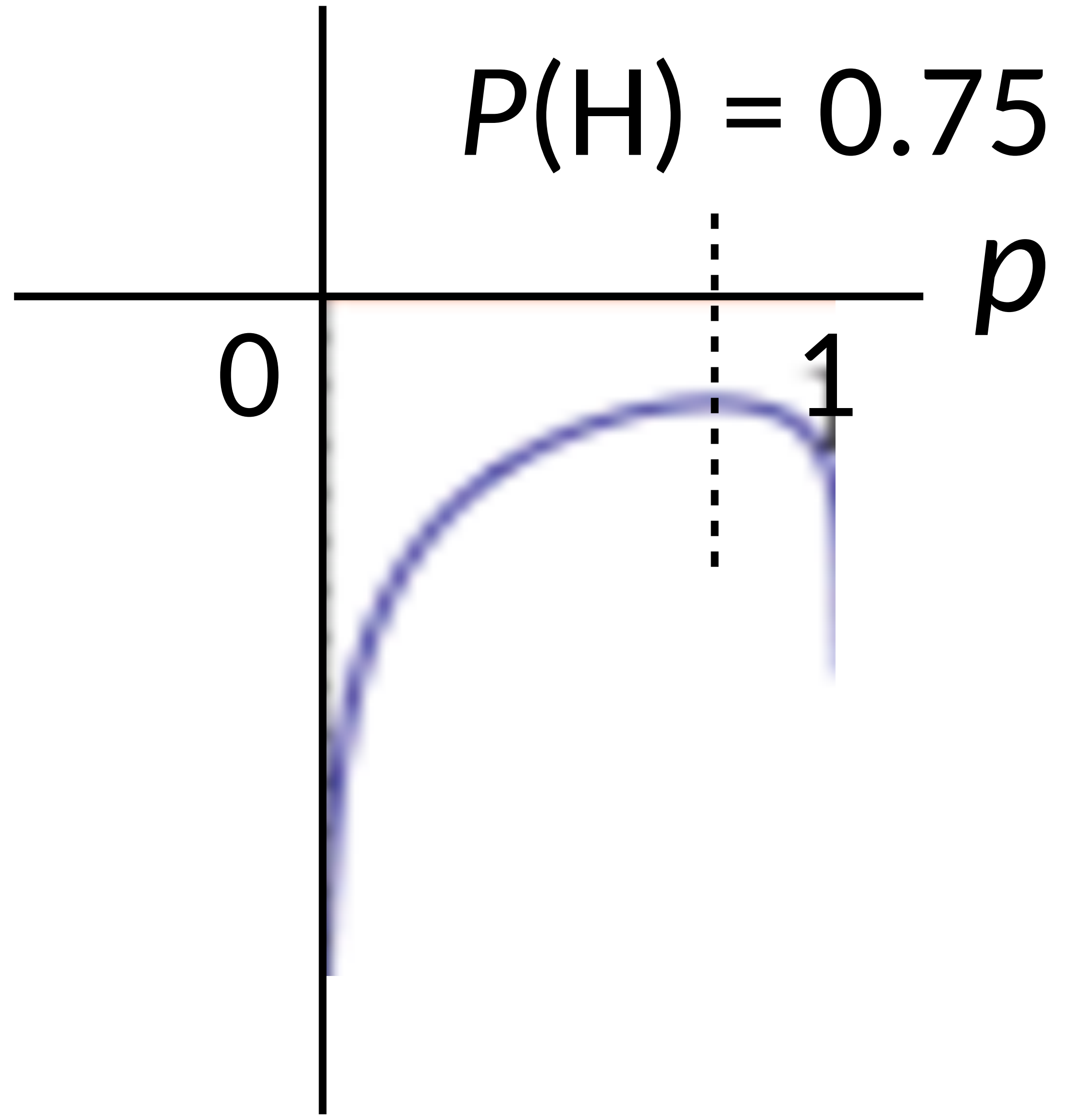
data points (*j*) features (*i*) *i*th feature of *j*th example

Maximum Likelihood Estimation

- Imagine a coin flip which is heads with probability *p*
- Observe (H, H, H, T) and maximize likelihood: $\prod_{j=1}^{m} P(y_j) = p^3(1-p)$
- Easier: maximize *log* likelihood

$$\sum_{j=1}^{m} \log P(y_j) = 3\log p + \log(1-p)$$

log likelihood



- Maximum likelihood parameters for binomial/ multinomial = read counts off of the data + normalize

Maximum Likelihood Estimation

- Data points (x_j, y_j) provided (*j* indexes over examples)
- Find values of $P(y), \;\; P(x_i|y)$ that maximize data likelihood (generative):

$$\prod_{j=1}^{m} P(y_j, x_j) = \prod_{j=1}^{m} P(y_j) \left[\prod_{i=1}^{n} P(x_{ji}|y_j) \right]$$

data points (*j*) features (*i*) *i*th feature of *j*th example

- Equivalent to maximizing logarithm of data likelihood:

$$\sum_{j=1}^{m} \log P(y_j, x_j) = \sum_{j=1}^{m} \left[\log P(y_j) + \sum_{i=1}^{n} \log P(x_{ji}|y_j) \right]$$

- Can do this by counting and normalizing distributions!

Summary

- You've now seen everything you need to implement multi-class classification models

- Homework (on Binary Classification): 09/09/2020 DUE

- Next: HMMs / POS tagging + CRFs (NER)