Text Classification

Jacob Andreas / MIT 6.806-6.864 / Spring 2020

Spam classification

RE: Staff and Faculty Mailbox Message !

Simona Matis <Simona.Matis@primariaclujnapoca.ro> Tue, Jan 14, 2020 at 9:20 AM
To: Simona Matis <Simona.Matis@primariaclujnapoca.ro>

Staff and Faculty Mailbox Message !

495MB **500**MB

This is to notify all Faculty Members and Staff on the end of year Mailbox Quota Cleanup, If you are a staff or faculty member log on to your staff and faculty ACCESS-PAGE **to clean up mailbox.**

Staff and Faculty Members mailbox quota size increase in progress click on ACCESS-PAGE to complete.

Mailbox Sending/Receiving authentication will be disabled at 500MB

ITS help desk
ADMIN TEAM

©Copyright 2020 Microsoft

[Supercloud-users] Reminder: Downtime

Lauren E Milechin <la25321@mit.edu> Wed, Jan 15, 2020 at 4:51 PM
Reply-To: supercloud@mit.edu
To: supercloud-users <supercloud-users@mit.edu>

Hello All,

Supercloud will be having its regular monthly downtime this Thursday, January 16 starting at midnight and ending about 24 hours later. We will send out an email when the downtime is complete the system is ready for jobs. The system may not come back up as quickly as it normally does, we are planning a few major changes. These changes should not change how you use the system.

As a reminder, we have a downtime scheduled every month, planned for the third Thursday of the month. We will continue to send reminder emails when the occur.

A general reminder: If you have any questions, please email supercloud@mit.edu.

Lauren

Spam classification

RE: Staff and Faculty Mailbox Message !

Simona Matis <Simona.Matis@primariaclujnapoca.ro> Tue, Jan 14, 2020 at 9:20 AM
To: Simona Matis <Simona.Matis@primariaclujnapoca.ro>

Staff and Faculty Mailbox Message !

495MB **500**MB

This is to notify all Faculty Members and Staff on ox Quota Cleanup, If you are a staff or faculty member log on to yo CCESS-PAGE to clean up mailbox.

Staff and Faculty Members mailbox quota size ACCESS-PAGE to complete.

Mailbox Sending/Receiving auth disabled at 500MB

ITS help desk
ADMIN TEAM

©Copyright 2020 Microsoft

spam

[Supercloud-users] Reminder: Downtime

Lauren E Milechin <la25321@mit.edu> Wed, Jan 15, 2020 at 4:51 PM
Reply-To: supercloud@mit.edu
To: supercloud-users <supercloud-users@mit.edu>

Hello All,

Supercloud will be having its regular monthly downtime this Thursday, Januar arting at midnight and ending about 24 hours later. We will send out an email when e is complete the system is ready for jobs. The system may not come back up as qu does, we are planning a few major changes. These changes should not cha ystem.

As a reminder, we have a downtime scheduled ever third Thursday of the month. We will continue to send reminder emails

A general reminder: If you have any questions, p supercloud@mit.edu.

Lauren

not spam

Аяны замд түр зогсон тэнгэрийн байдлыг ажиглаад хөдлөх зуур гутал дор шинэхэн орсон цас шаржигнан дуугарч байв. Цасны тухай бодол сонин юм. Хот хүрээ тийш цас орвол орно л биз гэсэн хэнэггүй бодол маань хөдөө талд, говийн ээрэм хөндийд, малын бэлчээрт, малчдын хотонд болохоор солигдож эргэцүүлэн бодох нь хачин. Цас хэр орсон бол?

Београд, 16. јун 2013. године – Председник Владе Републике Србије Ивица Дачић честитао је кајакашици златне медаље у олимпијској дисциплини К-1, 500 метара, као и у двоструко дужој стази освојене на првенству Европе у Португалији.

Language ID

Beograd, 16. jun 2013. godine – Predsednik Vlade Republike Srbije Ivica Dačić čestitao je kajakašici zlatne medalje u olimpijskoj disciplini K-1, 500 metara, kao i u dvostruko dužoj stazi osvojene na prvenstvu Evrope u Portugaliji.

Nestrankarski Urad za vladno odgovornost ZDA je objavil eksplozivno mnenje, da je vlada predsednika Donalda Trumpa kršila zvezno zakonodajo, ko je zadrževala izplačilo kongresno potrjene vojaške pomoči Ukrajini zaradi političnih razlogov. Predstavniški dom kongresa je prav zaradi tega sprožil ustavno obtožbo proti Trumpu.

Аяны замд түр зогсон тэнгэрийн байдлыг ажиглаад хөдлөх зуур гутал дор шинэхэн орсон цас шаржигнан дуугарч байв. Цасны тухай бодол сонин юм. Хот[redacted]рвол орно л биз гэсэн хэнэггүй бодол маань хөдөө тал[redacted]өндийд, малын бэлчээрт, малчдын хотонд болохоор солигдож эргэцүүлэн бодох нь хачин. Цас хэр орсон бол?

mongolian

Београд, 16. јун 2013. године – Председник Владе Републике Срб[redacted]неститао је кајакашици златне медаље у оли[redacted]ини К-1, 500 метара, као и у двоструко дужој стази освојене на првенству Европе у Португалији.

serbian

Language ID

Beograd, 16. jun 2013. godine – Predsednik Vlade Republike S[redacted]stitao je kajakašici zlatne medalje u o[redacted]i K-1, 500 metara, kao i u dvostruko dužoj stazi osvojene na prvenstvu Evrope u Portugaliji.

serbian

Nestrankarski Urad za vladno odgovornost ZDA je objavil eksplozivno mnenje, da je vlada predsednika Donalda Trumpa kršila zvezno zakonodajo, ko je zadrževala izplačilo k[redacted]vojaške pomoči Ukrajini zaradi političnih razlogov. Predstavniški d[redacted]av zaradi tega sprožil ustavno obtožbo proti Trumpu.

slovenian

By **John Neal**

This review is from: **Accoutrements Horse Head Mask (Toy)**

When I turned State's Witness, they didn't have enough money to put me in the Witness Protection Program, so they bought me this mask and gave me a list of suggested places to move. Since then I've lived my life in peace and safety knowing that my old identity is forever obscured by this life-saving item.

By **Christine E. Torok**

Verified Purchase (What's this?)

First of all, for taste I would rate these a 5. So good. Soft, true-to-taste fruit flavors like the sugar variety...I was a happy camper.

BUT (or should I say BUTT), not long after eating about 20 of these all hell broke loose. I had a gastrointestinal experience like nothing I've ever imagined. Cramps, sweating, bloating beyond my worst nightmare. I've had food poisoning from some bad shellfish and that was almost like a skip in the park compared to what was going on inside

By **John Neal**

This review is from: Accoutrements Horse Head Mask (Toy)

★★★★★

When I turned State's Witness, they didn't have enough money to put me in the Witness Protection Program, so they bought me this mask and gave me a list of suggested places to move. Since then I've lived my life in peace and safety knowing that my old identity is forever obscured by this life-saving item.

By **Christine E. Torok**

Verified Purchase (What's this?)

First of all, for taste I would rate these a 5. So good. Soft, true-to-taste fruit flavors like the sugar variety...I was a happy camper.

BUT (or should I say BUTT), not long after eating about 20 of these all hell broke loose. I had a gastrointestinal experience like nothing I've ever imagined. Cramps, sweating, bloating beyond my worst nightmare. I've had food poisoning from some bad shellfish and that was almost like a skip in the park compared to what was going on inside

★

Text classification

Input: document $x = [x_1, x_2, \ldots, x_n]$

[*gastrointestinal*, *experience*, ...]

[p, r, e, d, s, e d, n, ...]

Output: label $y \in \{1,2,3,\ldots,k\}$

`1 star`

`Serbian`

Rule-based classification

Rule-based classification

```
def classify_sentiment(document):
  for word in document:
    if word in {“good”, “wonderful”, “excellent”}:
      return 5
    if word in {“bad”, “awful”, “terrible”}:
      return 1
```

Rule-based classification: challenges

Sentiment: *Half submarine flick, half ghost story, all in one a criminally neglected film.*

→ hard to identify *a priori* which words are informative (and what information they carry!)

Rule-based classification: challenges

Sentiment: *Half submarine flick, half ghost story, all in one a criminally neglected film.*

→ hard to identify *a priori* which words are informative (and what information they carry!)

Sentiment: *It's not life-affirming, it's vulgar, it's mean, but I liked it.* → word order matters, but hard to encode in rules!

Rule-based classification: challenges

Sentiment: *Half submarine flick, half ghost story, all in one a criminally neglected film.*

→ hard to identify *a priori* which words are informative (and what information they carry!)

Sentiment: *It's not life-affirming, it's vulgar, it's mean, but I liked it.* → word order matters, but hard to encode in rules!

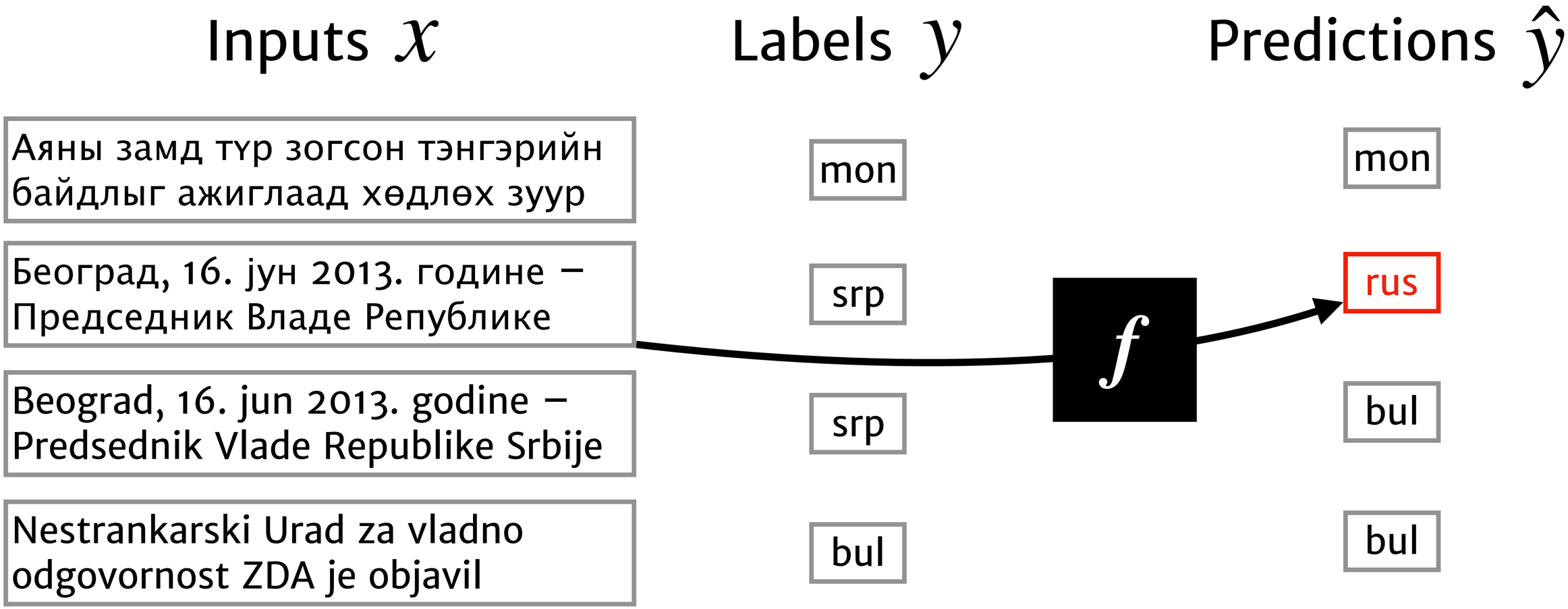
Language ID: *All falter, stricken in kind.* *"LINGERIE SALE"*

→ simple features can be misleading!

Rule-based classification

But don’t forget: if you don’t have access to data, speaker intuition and a bit of coding get you pretty far!

Learning-based classification



Goal: pick the function *f* that does “best” on training data

Where do datasets come from?

Human institutions

Noisy labels

Expert annotation

Crowd workers

Where do datasets come from?

Human institutions

Government proceedings

Product reviews

Noisy labels

Expert annotation

Crowd workers

Where do datasets come from?

Human institutions	Noisy labels	Expert annotation	Crowd workers
Government proceedings	Domain names		
Product reviews	Link text		

Where do datasets come from?

Human institutions	Noisy labels	Expert annotation	Crowd workers
Government proceedings	Domain names	Treebanks	
Product reviews	Link text	Biomedical corpora	

Where do datasets come from?

Human institutions	Noisy labels	Expert annotation	Crowd workers
Government proceedings	Domain names	Treebanks	Question answering
Product reviews	Link text	Biomedical corpora	Image captions

Linear classifiers

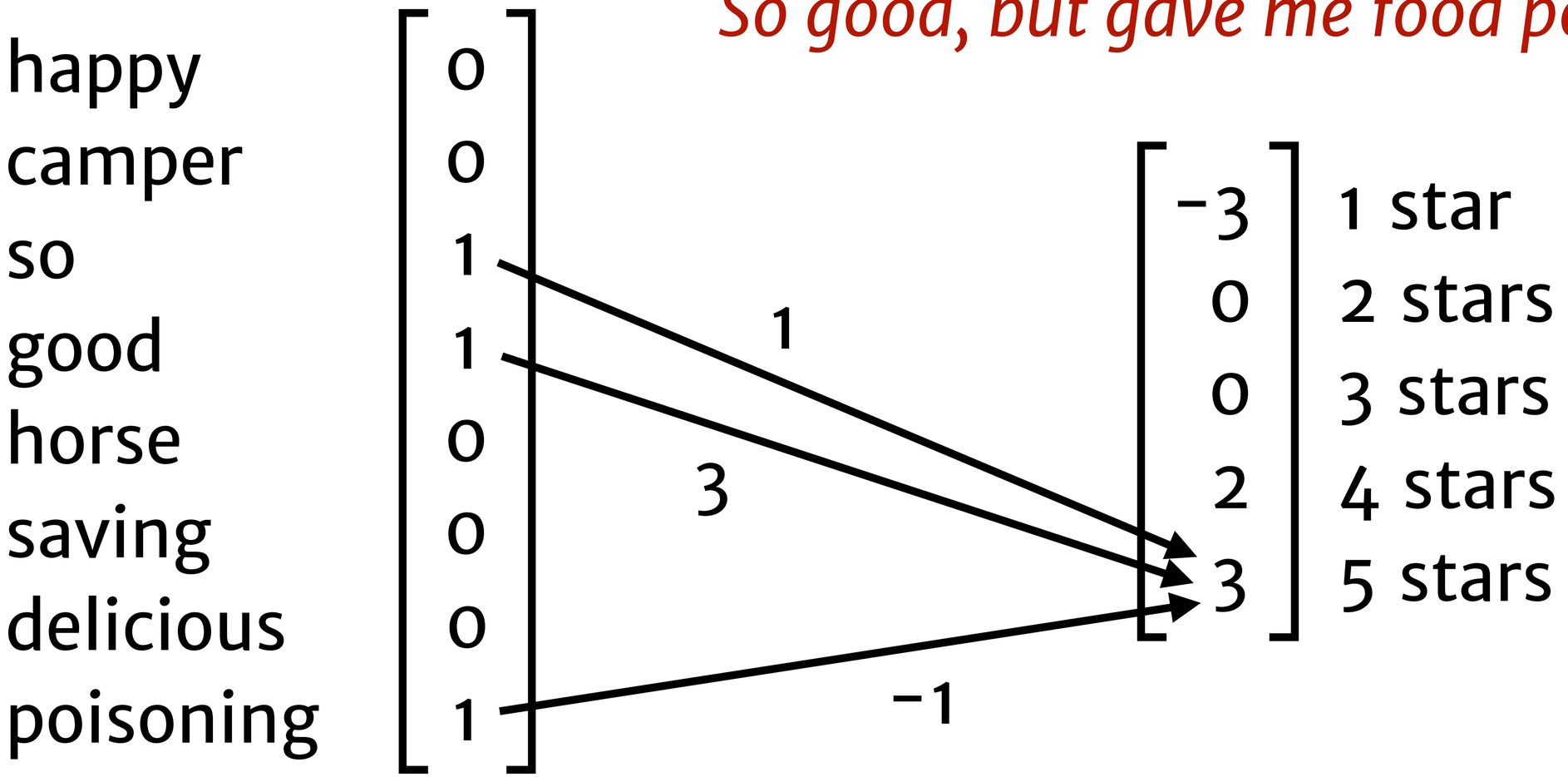
Linear models

happy	0
camper	0
so	1
good	1
horse	0
saving	0
delicious	0
poisoning	1

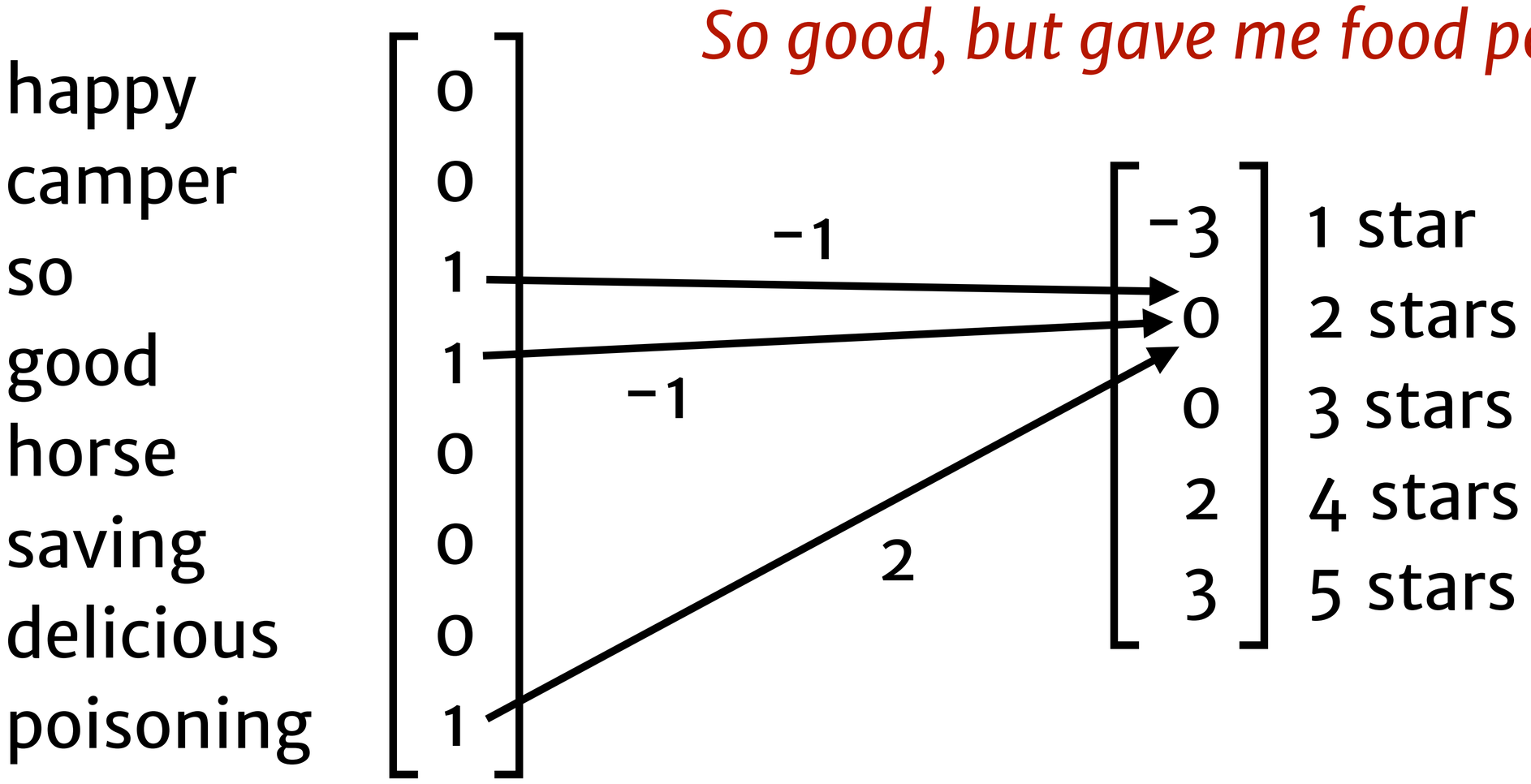
So good, but gave me food poisoning.

Linear models

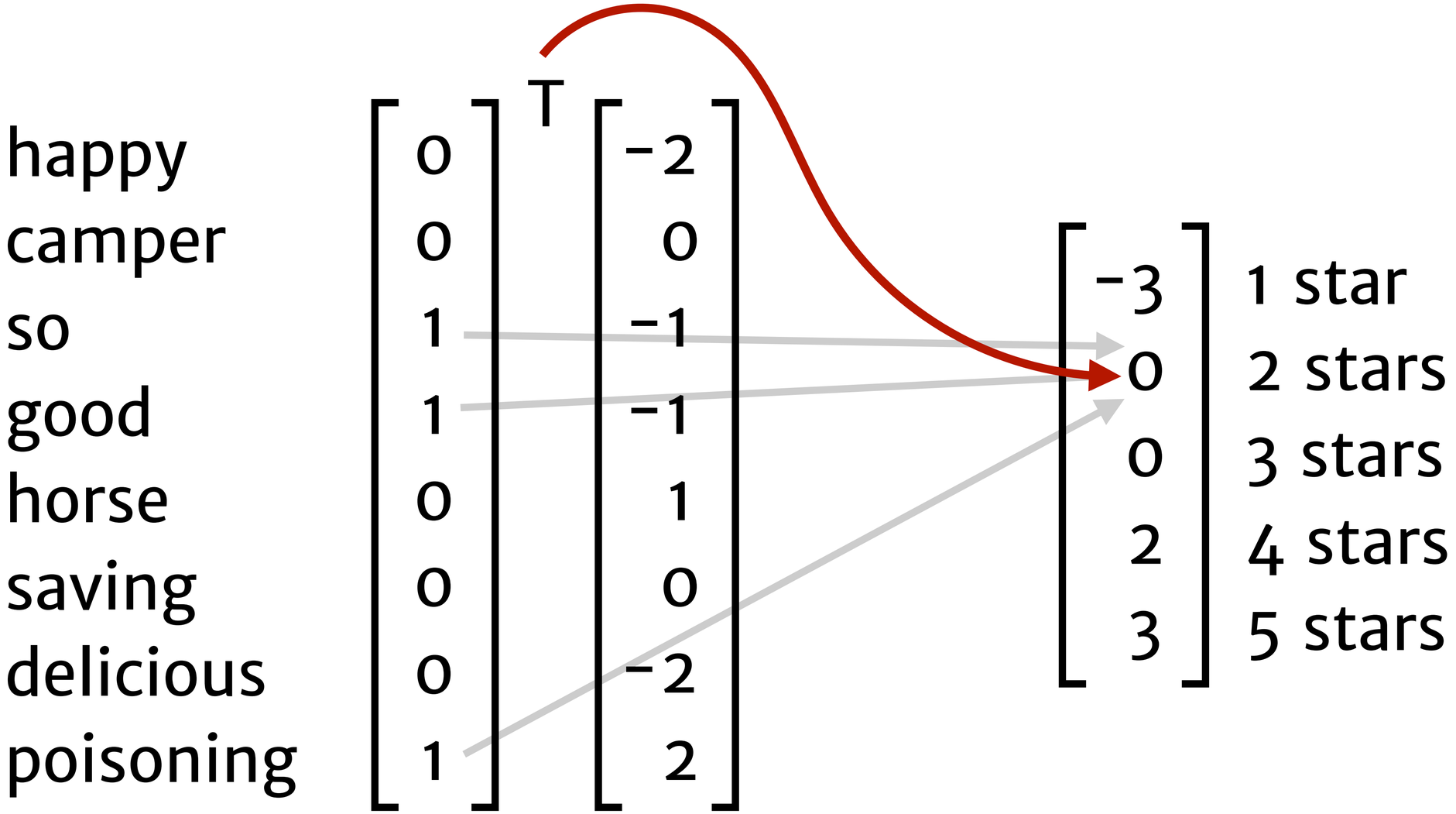
So good, but gave me food poisoning.



Linear models



Linear models



Linear models

scores weights

$$S = W^\top x$$

features

$n_{classes}$

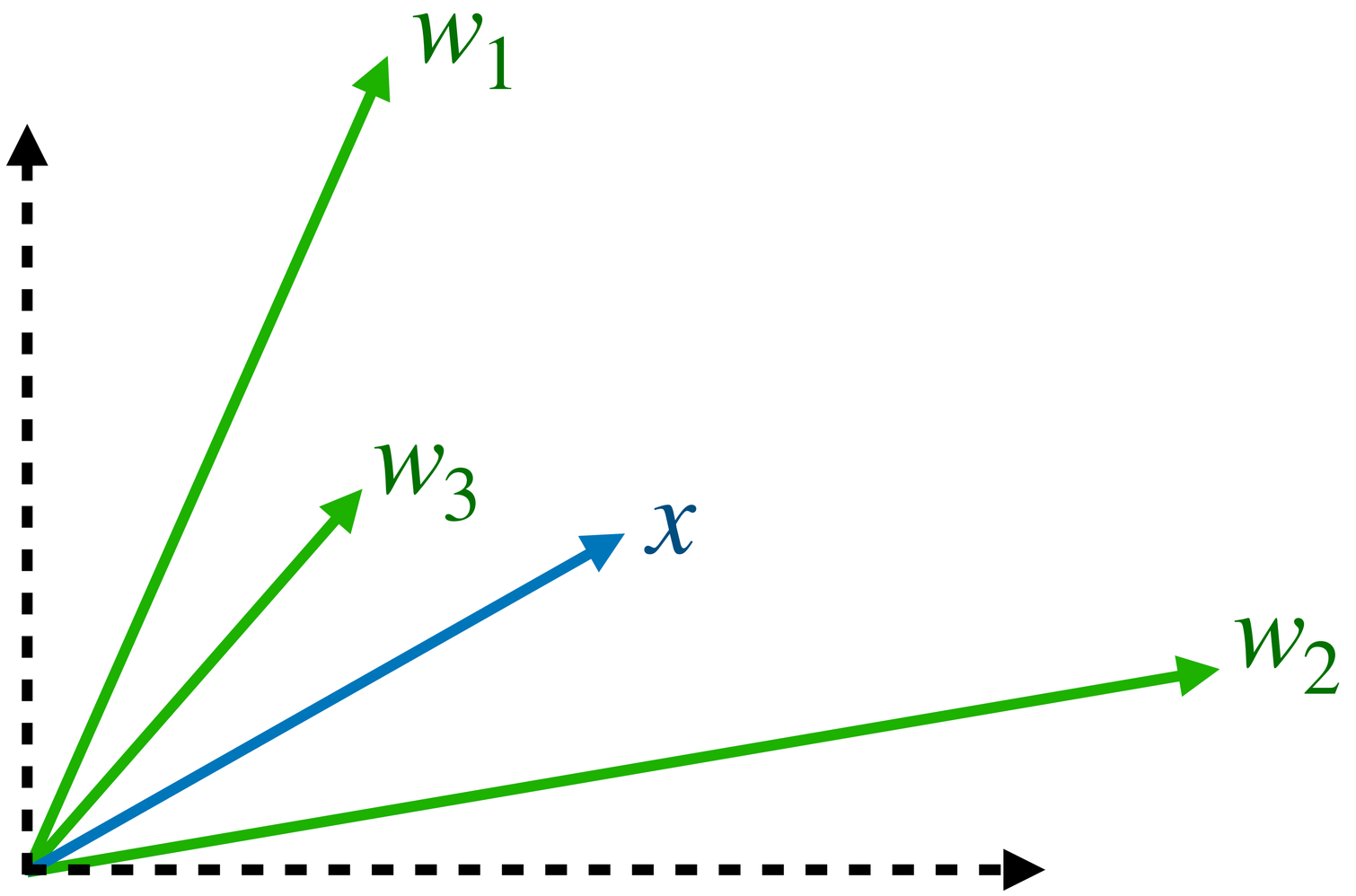
$n_{features}$

$n_{features}$ x $n_{classes}$

Linear models

$$s = W^\top x$$

$$s_i = w_i^\top x$$



Feature design

Not the worst eraser I've ever eaten.

Feature design

Not the worst eraser I've ever eaten.

words

not

worst

eaten

eraser

Feature design

Not the worst eraser I've ever eaten.

words	skip-grams
not	*not the*
worst	*worst eraser*
eaten	*not…worst*
eraser	

Feature design

Not the worst eraser I've ever eaten.

words	skip-grams	structural features
not	*not the*	`length=7`
worst	*worst eraser*	`%capitalized=14`
eaten	*not…worst*	
eraser		

Feature design

Not the worst eraser I've ever eaten.

words	skip-grams	structural features	products
not	*not the*	`length=7`	*worst* & `length=7`
worst	*worst eraser*	`%capitalized=14`	*not the* & *worst*
eaten	*not…worst*		
eraser			

Learning

How well does $W^\top x = s$ predict y?

Define a loss function L, and choose W to minimize loss across data points:

$$\text{argmin}_W \sum_{(x,y)} L(s, y)$$

Learning: exact match

$$L(s, y) = \mathbf{1}[\text{argmax}(s) = y]$$

$$\text{argmin}_W \sum_{(x,y)} L(s, y)$$

Hard to find a good W!

Learning: negative log likelihood

$$L(s, y) = -\log p(y \mid x)$$

$$= -\log \frac{\exp(s_y)}{\sum_i \exp(s_i)}$$

$$= -s_y + \log \sum_i \exp(s_i) := -\log \text{softmax}(\text{s})_y$$

Idea: treat s as a vector of (unnormalized) log-probs, and maximize p(y | x; W).

Learning: margin

$$L(s, y) = [\max(s_{-y}) - s_y + c]_+$$

s_{-y}: scores other than y

$[x]_+ := \max(x, 0)$

Idea: try to make the score of the right label s_y at least at least c greater than the score of every wrong label.

Learning: optimization & regularization

$$\text{argmin}_W \sum_{(x,y)} L(s, y)$$

How to compute W? SGD!

$$W_{t+1} = W_t - \alpha \nabla L(s, y)$$

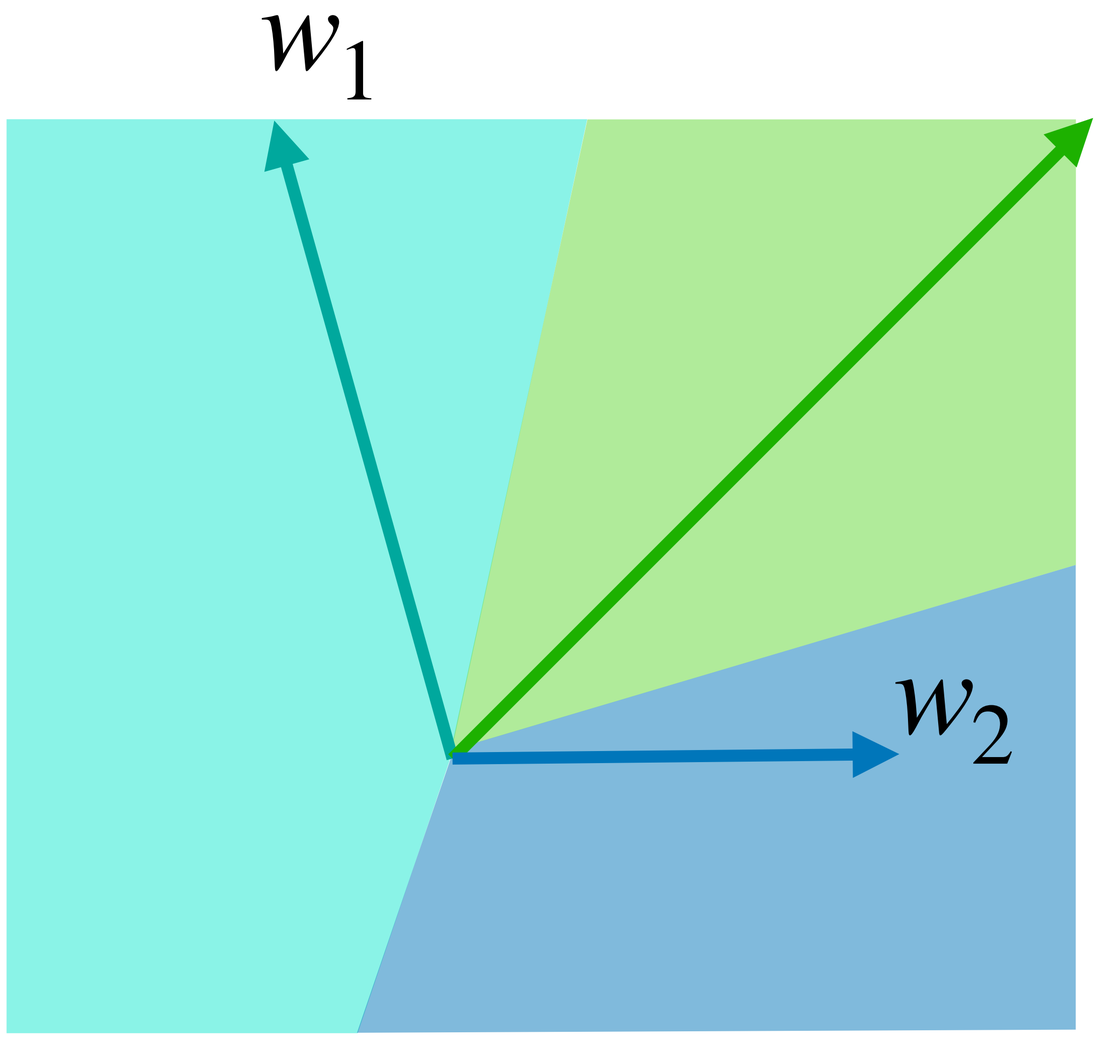
Inference and uncertainty

Given a new x, what y should we predict?

$$\hat{y} = \text{argmax}_i \; s_i$$

If we trained with a likelihood loss, we can also interpret $\text{softmax}(s)$ as a prediction of $p(y \mid x)$. (What about a margin loss?)

Linear decision boundaries



$$\hat{y} = \text{argmax}_i \; w_i^\top x$$

Overfitting

complex features

$$\text{argmin}_W \sum_{(x,y)} L(s, y)$$

worst &
`length=7`

yesterday…five &
length=26

This is not a very useful feature, but it probably occurs just once in the training set. So it's easy to "memorize" the label in a way that won't generalize.

Regularization

Introduce a regularizer $R(W)$ that encourages "simple" solutions:

$$\text{argmin}_W \left[\sum_{(x,y)} L(s, y) \right] + R(W)$$

ℓ^2 regularization

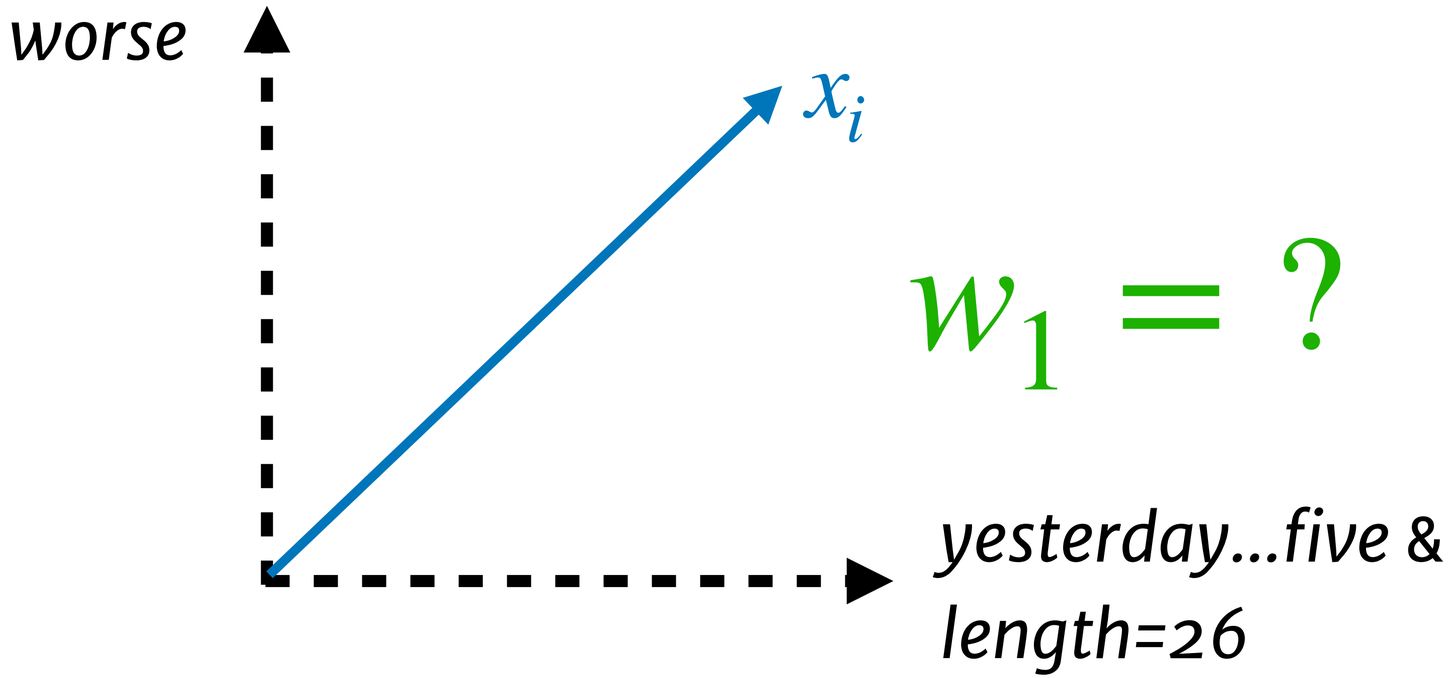
$$R(W) = \|W\|_2^2 = \sum_i W_i^2$$

ℓ^2 regularization

$$R(W) = \|W\|_2^2 = \sum_i W_i^2$$

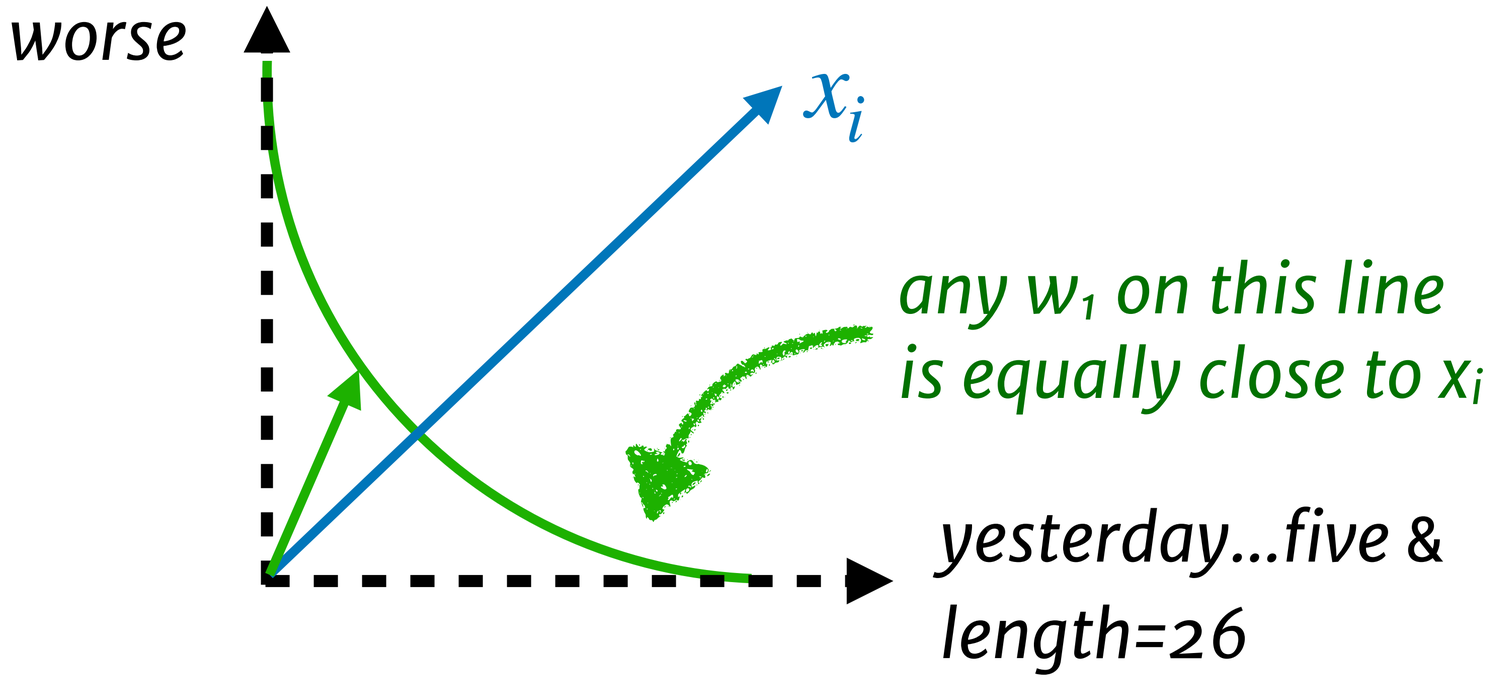
$x =$ *the movie I saw yesterday was five times worse than the worst movie...*

$y = 1$



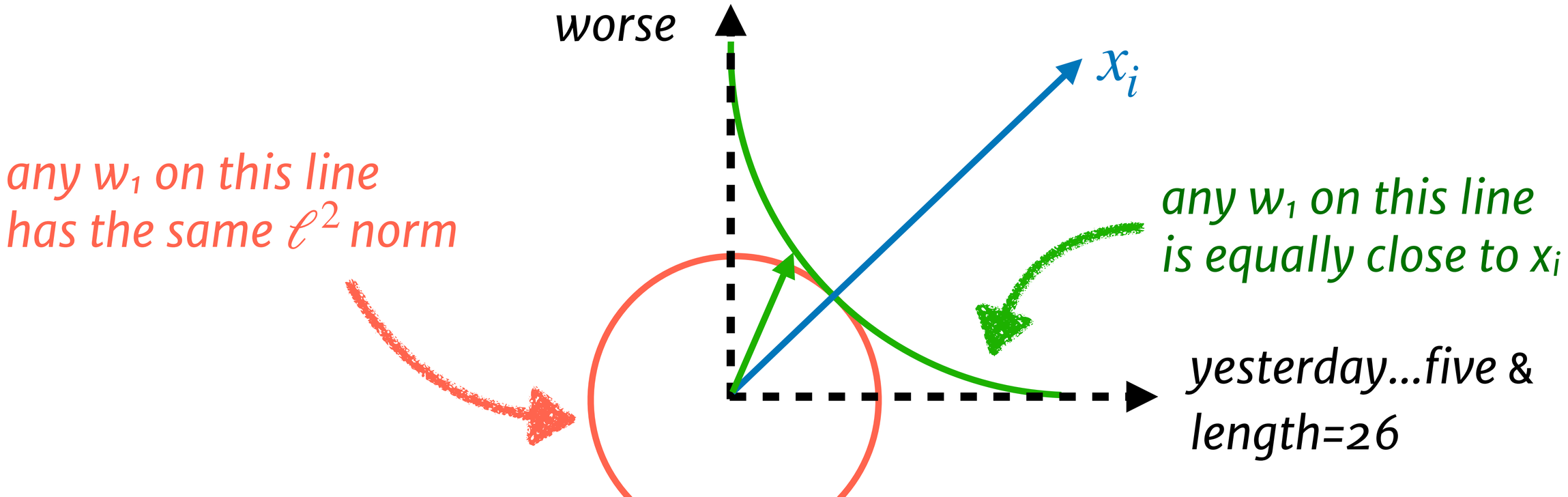
ℓ^2 regularization

$$R(W) = \|W\|_2^2 = \sum_i W_i^2$$



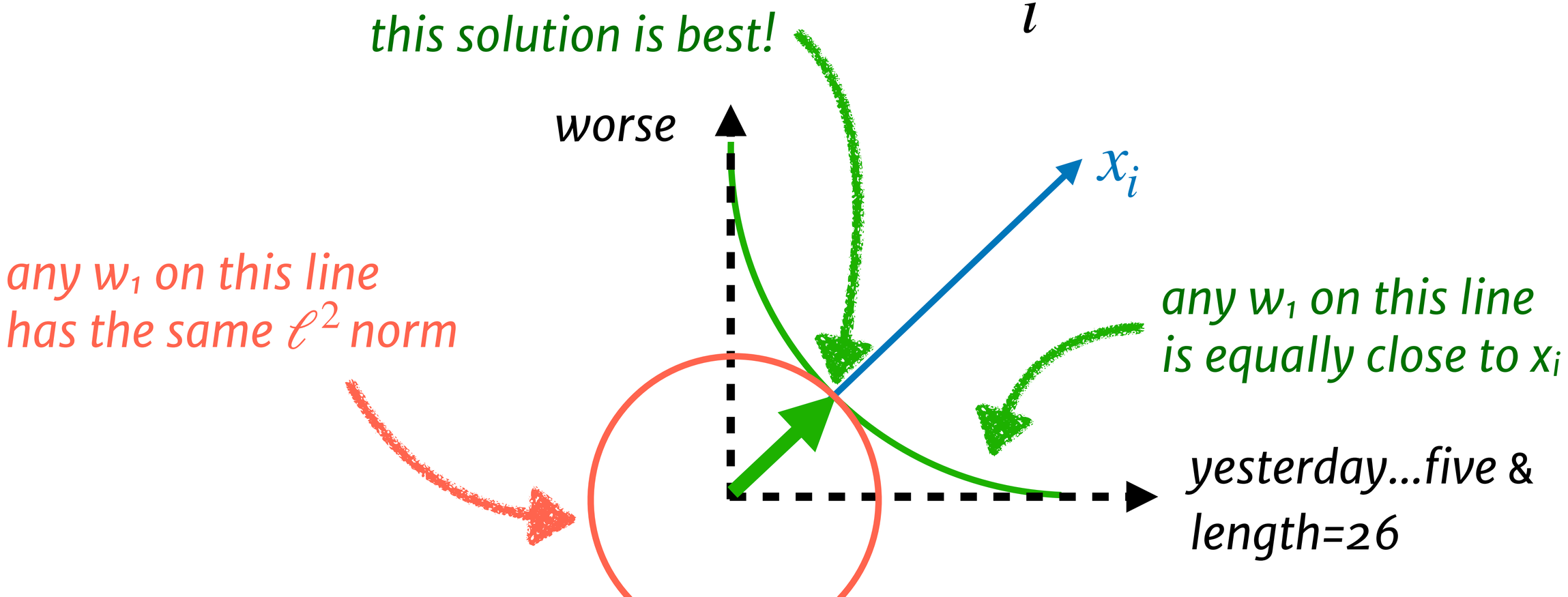
ℓ^2 regularization

$$R(W) = \|W\|_2^2 = \sum_i W_i^2$$



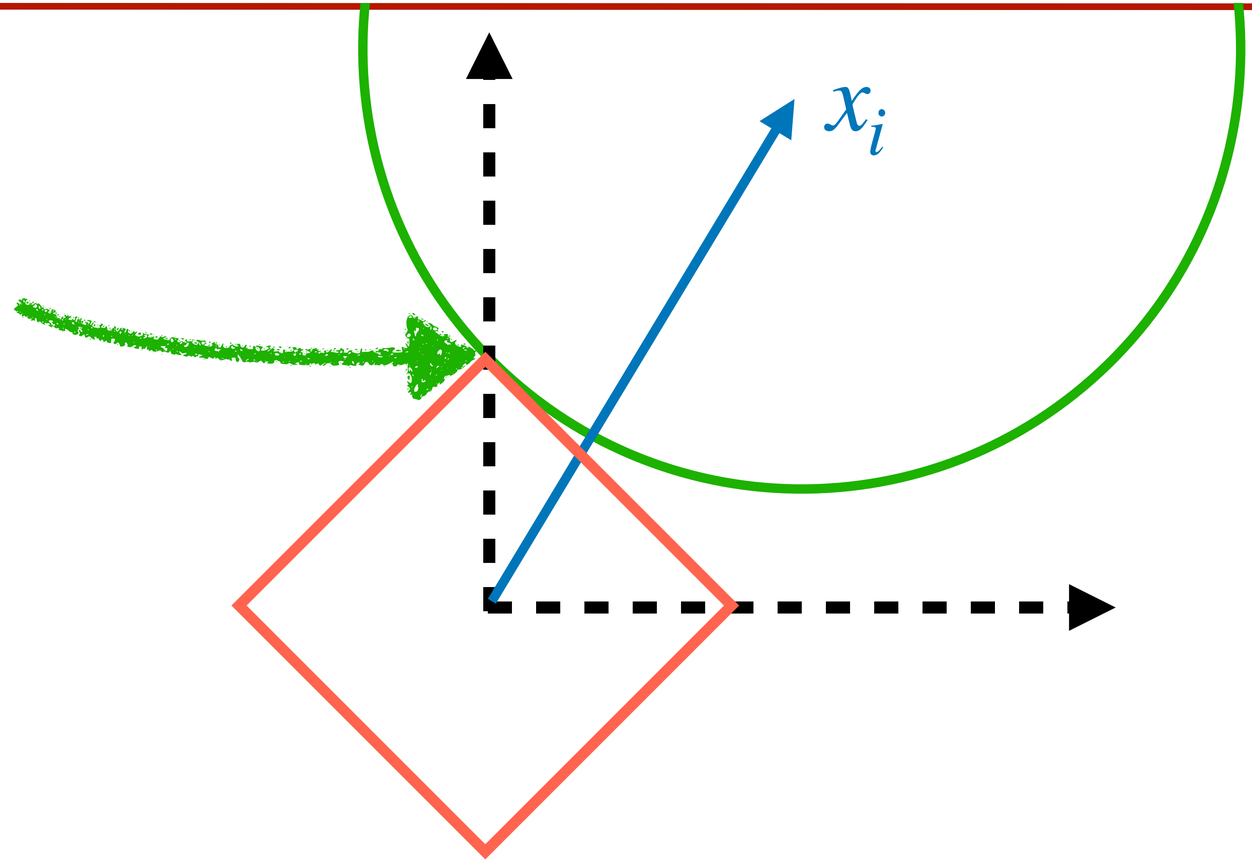
ℓ^2 regularization

$$R(W) = \|W\|_2^2 = \sum_i W_i^2$$



ℓ^1 regularization

this solution is best:
sparsity!



$$R(W) = \|W\|_1 = \sum_i |W_i|$$

Overfitting

complex features

$$\text{argmin}_W \sum_{(x,y)} L(s, y) + R(W)$$

worst &
`length=7`

yesterday…five &
length=26

← With regularization, it's generally safe to include large numbers of irrelevant features.

Challenges: data sparsity

$x =$ *perhaps not the most scintillating work in the director's oeuvre*

Challenges: data sparsity

$x =$ *perhaps not the most scintillating work in the director's oeuvre*

good	5	
oeuvre	0	???
terrible	−4	

$$y = ?$$

Most people don't write like this, so we might not see *any* of the content words associated with labels.

Challenges: data sparsity[2]

$x =$ *perhaps not the most scintillating work in the director's oeuvre*

somewhat...awful	−1	???
not...scintillating	0	
less...terrible	1	

$$y = ?$$

Rare combinations of features will provide even less of a signal to learn from.

Challenges: nonlinear feature interactions

bad

Challenges: nonlinear feature interactions

bad

not bad

Challenges: nonlinear feature interactions

bad

not bad

not so bad

Challenges: nonlinear feature interactions

bad

not bad

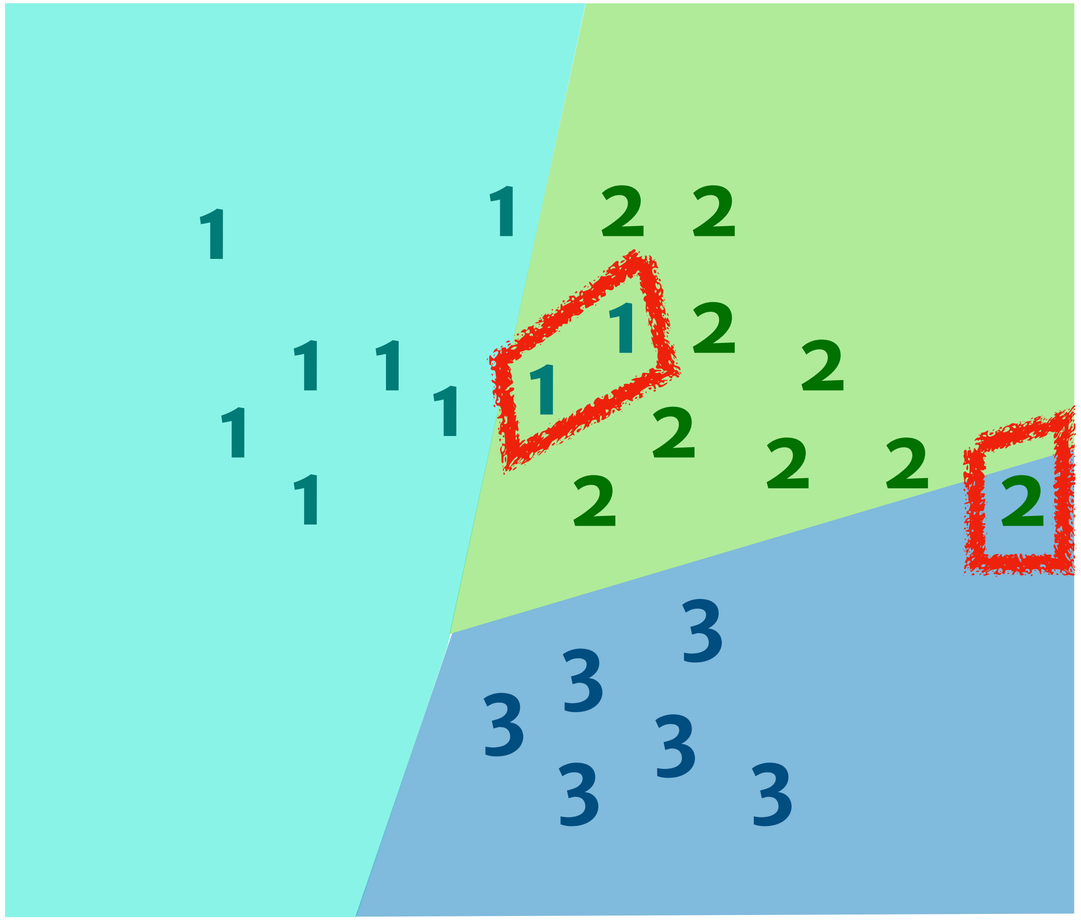
not so bad

wouldn't have been so bad but

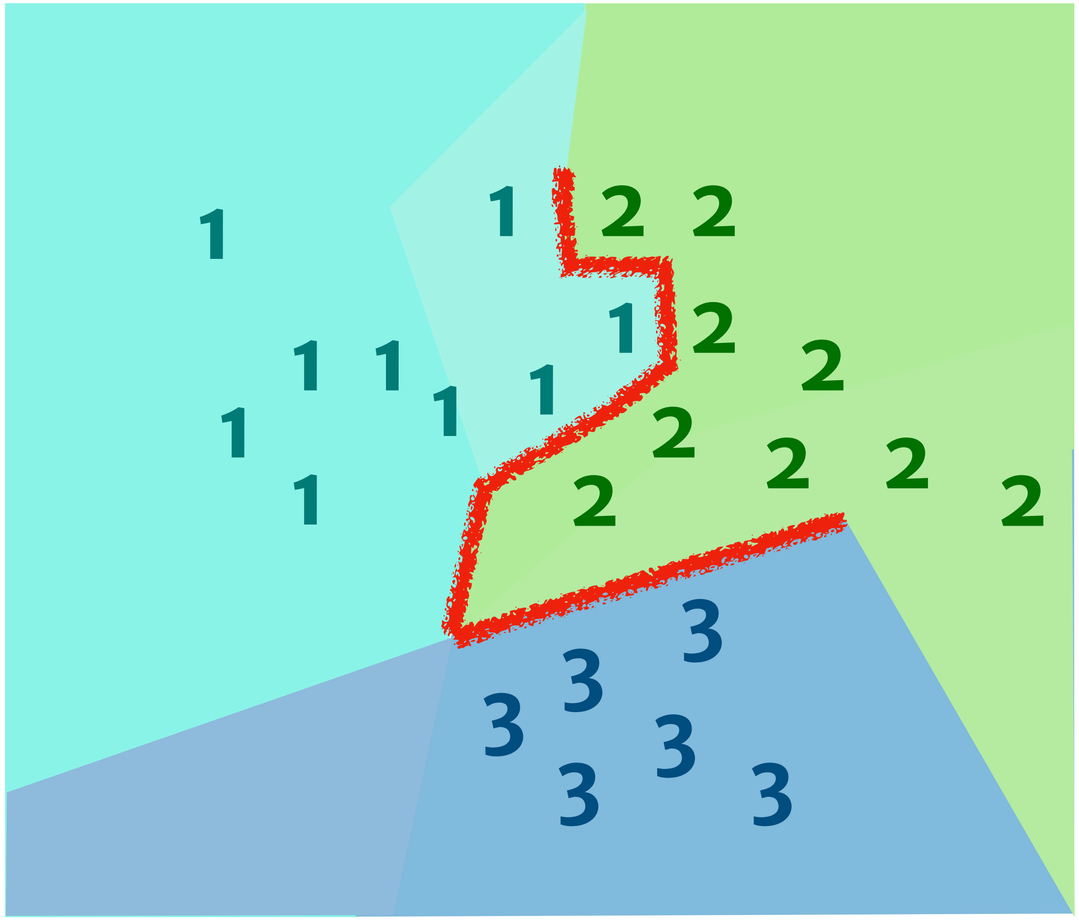
Not all feature interactions can be captured by adding product terms.

Nearest-neighbor classifiers

Nearest-neighbor models

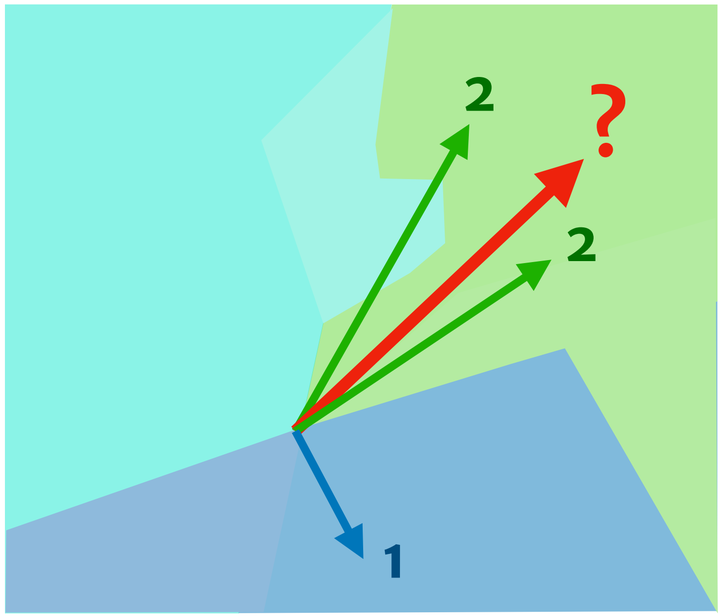


Nearest-neighbor models



Model formulation

$$s(x)_y := \sum_{(x',y'):\ y'=y} \text{similarity}(x,\ x')$$



nearest neighbors:

$\text{similarity}(x,\ x') =$ 1 if x' is one of the k closest points to x else 0

kernel methods:

$\text{similarity}(x,\ x') = e^{-||x-x'||}$ e.g.

Representation design

Not the worst eraser I've ever eaten.

n-grams	skip-grams	structural features	products
not	*not the*	`length=7`	*worst* & `length=7`
worst	*worst eraser*	`%capitalized=14`	*not the* & *worst*
eaten	*not…worst*		
eraser			

“Learning”

$$s(x)_y := \sum_{(x',y'):\ y'=y} \text{similarity}(x,\ x')$$

Just memorize the training dataset!

Inference and uncertainty

Given a new x, what y should we predict?

$$\hat{y} = \text{argmax}_i \; s_i$$

(Probabilistic interpretation?)

Neural networks

Deep network models

Linear model: $s = W^\top x$

Deep linear model: $s = W_2^\top W_1^\top$

(same expressive power!)

Neural network model: $s = W_2^\top f(W_1^\top x)$

Deep network models

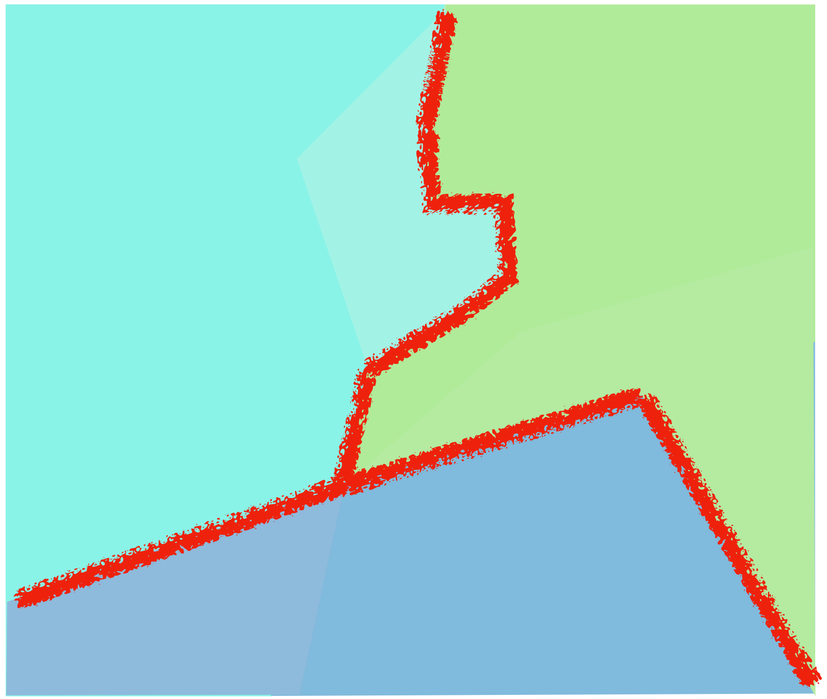
$$f(x) = \text{e.g. } [x]_+, \; 1/1+e^{-x}$$

"ReLU" "sigmoid"

Neural network model: $s = W_2^\top f(W_1^\top x)$

Multilayer perceptron

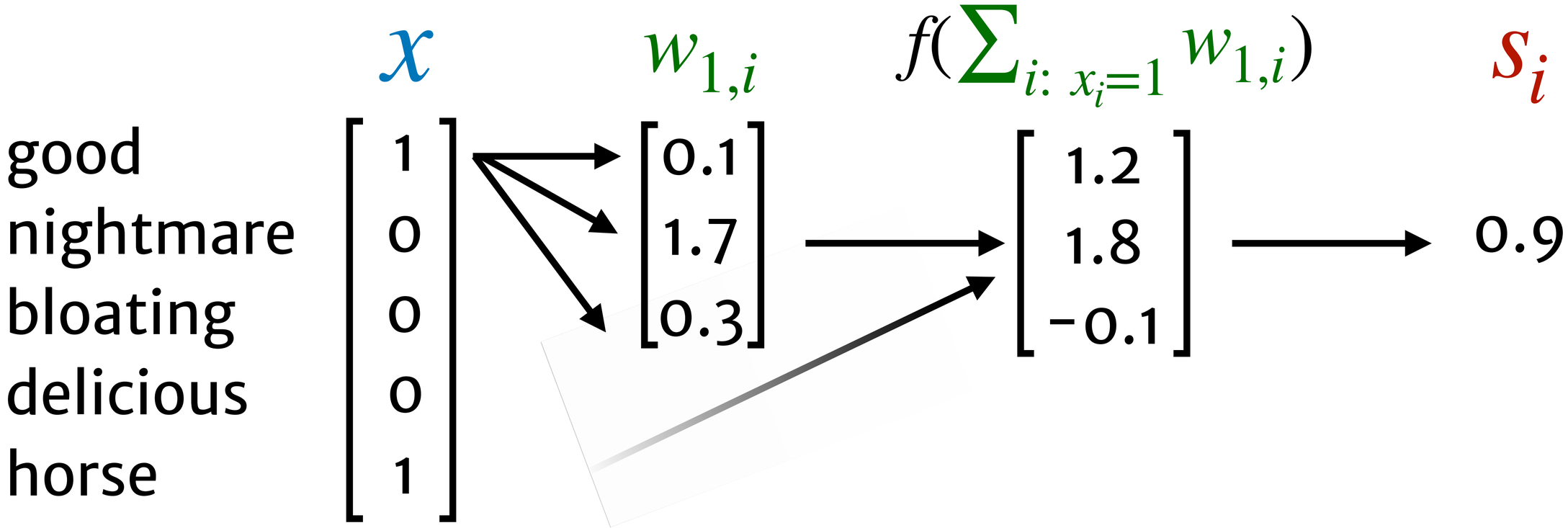
$$s = W_2^\top f(W_1^\top x)$$



Nonlinear decision boundaries!

Interpretation: deep bag of input features

$$s = W_2^\top f(W_1^\top x)$$



Interpretation: deep bag of words

$$s = W_2^\top f(W_1^\top x)$$

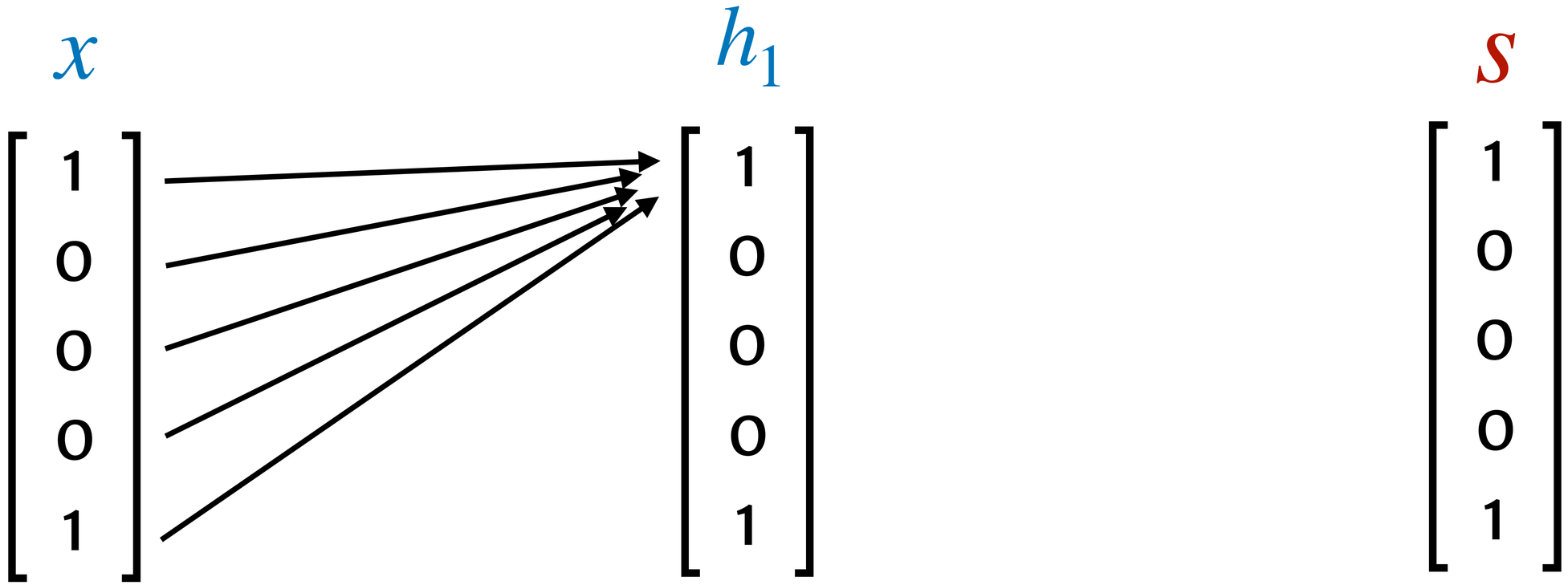
$w_{1,i}^\top$

$$\begin{bmatrix}0.1\\1.7\\0.3\end{bmatrix}\begin{bmatrix}0.1\\1.7\\0.3\end{bmatrix}\begin{bmatrix}0.1\\1.7\\0.3\end{bmatrix} \xrightarrow{f(\sum_{i:\ x_i=1} w_i)} \begin{bmatrix}1.2\\1.8\\-0.1\end{bmatrix} \longrightarrow 0.9$$

x *very good, but gave me*

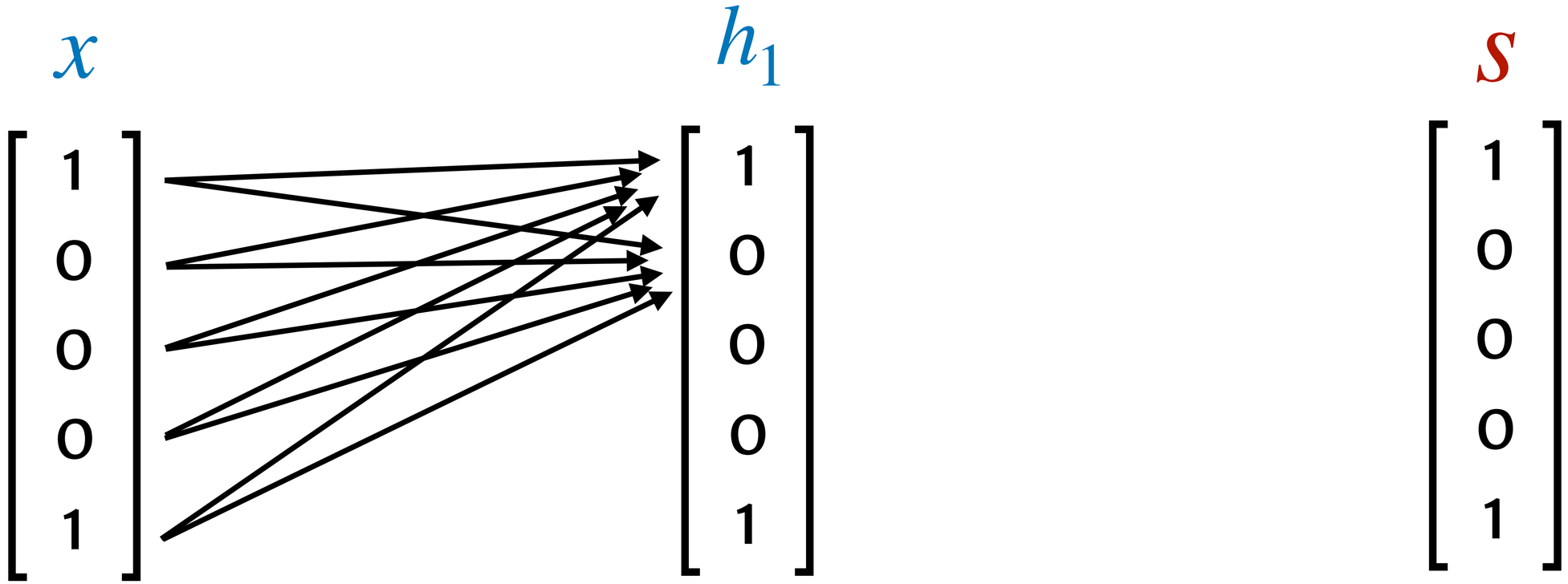
Interpretation: deep bag of words

$$s = W_2^\top f(W_1^\top x)$$



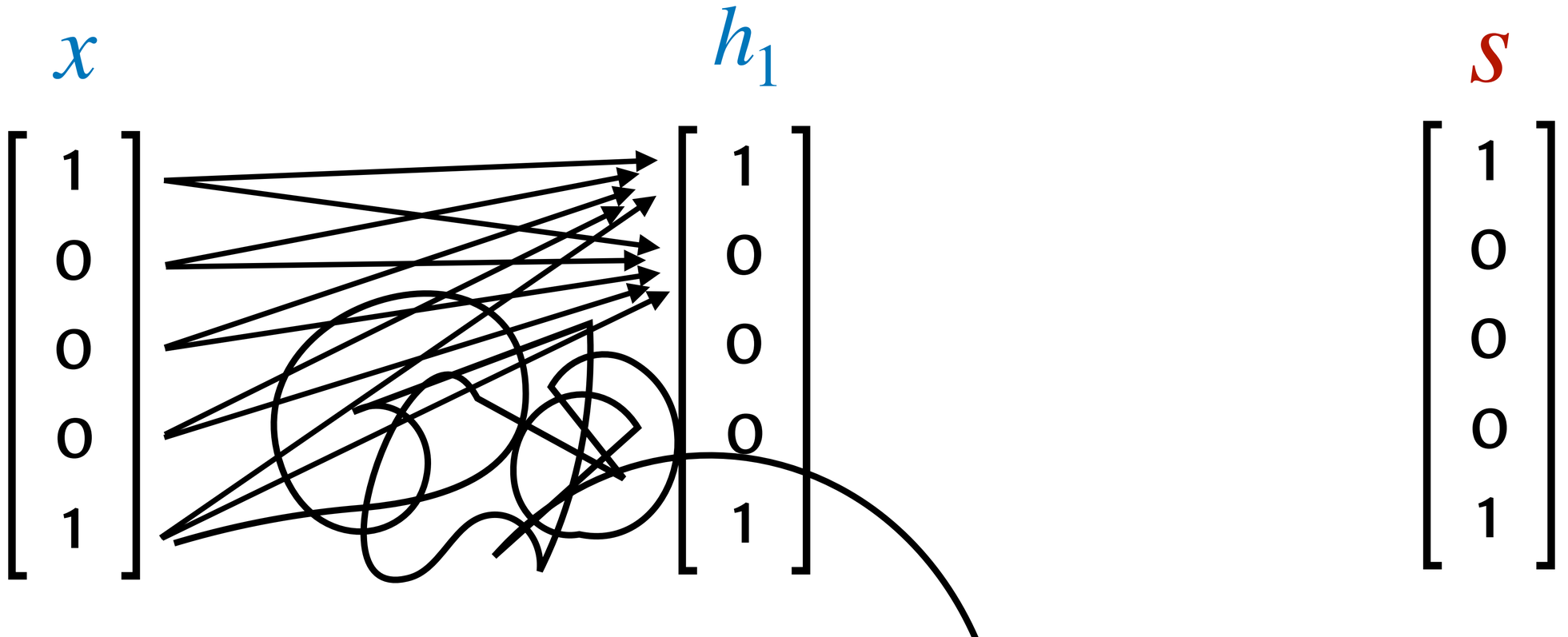
Interpretation: deep bag of words

$$s = W_2^\top f(W_1^\top x)$$



Interpretation: deep bag of words

$$s = W_2^\top f(W_1^\top x)$$



Interpretation: deep bag of words

$$s = W_2^\top f(W_1^\top x)$$

x

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

input

$f(W_1^\top x) = h_1$

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

“hidden layer”

$W_2^\top h_1 = s$

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

output

Representation design revisited

Not the worst eraser I've ever eaten.

words	skip-grams	structural features	products
not	*not the*	`length=7`	*worst* & `length=7`
worst	*worst eraser*	`%capitalized=14`	*not the* & *worst*
eaten	*not…worst*		
eraser	(???)		

Challenges

Much larger design space:

What nonlinearity to use?
(Common choices: ReLU, sigmoid, tanh)

What optimizer?

How many hidden layers? How big?

How to tokenize the input?

How to capture ordering information?

Text classification in the real world

In the real world

You are building tools that will affect people's lives!

Always ask:

Is this model reliable enough to deploy at all?

In what ways does my data reflect (or fail to reflect) the "real world?"

Will different groups experience disparate prediction qualities?

Example: bias

Feature weights from a restaurant review star classifier:

asian fusion	**-0.38**
chinese	**-0.21**
coffee	**0.30**
diner	**-0.45**
ethiopian	**0.10**
fast food	**-0.01**
french	**0.16**
greek	**0.56**

[Jurafsky et al., First Monday 2014]

Adversarial inputs

Harassment detection model:

you	**-0.1**
ugly	**2.3**
hate	**3.1**
kill	**5.1**
nasty	**1.2**
friends	**-0.2**

Adversarial inputs

You're ugly and everyone hates you.

score: 10.7, label: possible harassment

Harassment detection model:

you	**-0.1**
ugly	**2.3**
hate	**3.1**
kill	**5.1**
nasty	**1.2**
friends	**-0.2**

Adversarial inputs

Harassment detection model:

you	**-0.1**
ugly	**2.3**
hate	**3.1**
kill	**5.1**
nasty	**1.2**
friends	**-0.2**

You're ugly and everyone hates you.

score: 10.7, label: possible harassment

You're ugly, everyone hates you, and you have no friends.

score: 10.3, label: possible harassment

Adversarial inputs

Harassment detection model:

you	**-0.1**
ugly	**2.3**
hate	**3.1**
kill	**5.1**
nasty	**1.2**
friends	**-0.2**

You're ugly and everyone hates you.

score: 10.7, label: possible harassment

You're ugly, everyone hates you, and you have no friends.

score: 10.3, label: possible harassment

You're ugly, everyone hates you, and you have no friends.
you you friends the and Monday happy good !!??

score: -6.1, label: no harassment

Adversarial inputs

Weird inputs make simple models behave in weird ways!

You're ugly and everyone hates you.

score: 10.7, label: possible harassment

You're ugly, everyone hates you, and you have no friends.

score: 10.3, label: possible harassment

You're ugly, everyone hates you, and you have no friends.
you you friends the and Monday happy good !!??

score: -6.1, label: no harassment

Summary

Today: learning to predict categorical labels from text

Linear models: you can do a lot with simple features and matrix multiplication!

Neural networks: nonlinear feature combinations “for free”, less feature engineering but larger design space

Next class: distributional semantics