Probability Review

CSE 447 / 517
January 13th, 2022 (Week 2)

Eisenstein (2019) 6 and Appendix A

Logistics

- Assignment 1 deadline extended!

  Now due **TODAY(1/13) 11:59 PM PST**

Agenda

- Quiz 1 Solutions
- Probability Review
- Assignment 1 Q & A

Quiz 1 - Problem Recap

Imagine you are in charge of implementing a text classifier for Gmail to sort incoming emails into categories.

- Primary: Emails from people you know (and messages that don't appear in other tabs)
- Social: Messages from social networks and media-sharing sites
- Promotions: Deals, offers, newsletters and other "call to action" emails
- Updates: Notifications, confirmations, receipts, bills and statements
- Forums: Messages from online groups, discussion boards and mailing lists

To simplify the problem, assume you only have access to the subject of the email, and all of the emails are in English.

Your goal is to build a multinomial logistics regression text classifier and evaluate its performance.

Quiz 1 - Problem 1

One of the emails in our dataset reads: "AMAZING SALE! Black Friday sale: 30% Off!" ...

Suppose we apply term frequency features for the given email example, what would the value for the following feature be? Assume we lowercase everything, removed all symbols and punctuations, and split the text into tokens by white spaces.

$$\phi^{freq}_{amazing}(\mathbf{x}) =$$

$$\phi^{freq}_{awesome}(\mathbf{x}) =$$

$$\phi^{freq}_{big}(\mathbf{x}) =$$

$$\phi^{freq}_{friday}(\mathbf{x}) =$$

$$\phi^{freq}_{off}(\mathbf{x}) =$$

$$\phi^{freq}_{deal}(\mathbf{x}) =$$

$$\phi^{freq}_{sale}(\mathbf{x}) =$$

$$\phi^{freq}_{30}(\mathbf{x}) =$$

Quiz 1 - Problem 1

One of the emails in our dataset reads: "AMAZING SALE! Black Friday sale: 30% Off!" ...

Suppose we apply term frequency features for the given email example, what would the value for the following feature be? Assume we lowercase everything, removed all symbols and punctuations, and split the text into tokens by white spaces.

$$\phi_{amazing}^{freq}(\mathbf{x}) = 1$$
$$\phi_{awesome}^{freq}(\mathbf{x}) = 0$$
$$\phi_{big}^{freq}(\mathbf{x}) = 0$$
$$\phi_{friday}^{freq}(\mathbf{x}) = 1$$

$$\phi_{off}^{freq}(\mathbf{x}) = 1$$
$$\phi_{deal}^{freq}(\mathbf{x}) = 0$$
$$\phi_{sale}^{freq}(\mathbf{x}) = 2$$
$$\phi_{30}^{freq}(\mathbf{x}) = 1$$

Quiz 1 - Problem 2

Suppose we use the words in Question 1 as the set of words that we work with (V = {amazing, awesome, big, friday, off, deal, sale, 30}). How many features would there be for our multinomial logistics regression model?

Quiz 1 - Problem 2

Suppose we use the words in Question 1 as the set of words that we work with (V* = {amazing, awesome, big, friday, off, deal, sale, 30}). How many features would there be for our multinomial logistics regression model?

$$f^{\text{freq}}_{\text{Primary, amazing}}(\mathbf{x}, y)$$

$$f^{\text{freq}}_{\text{Primary, awesome}}(\mathbf{x}, y)$$

$$\ldots$$

$$f^{\text{freq}}_{\text{Primary, 30}}(\mathbf{x}, y)$$

$$f^{\text{freq}}_{\text{Social, amazing}}(\mathbf{x}, y)$$

$$\ldots$$

$$f^{\text{freq}}_{\text{Forums, 30}}(\mathbf{x}, y)$$

There 8 tokens in our vocabulary.

There are 5 different labels.

We need **40** features in total.

Quiz 1 - Problem 3

Which ones of the following might be other reasonable features to use for our problem? (enter "true" if it is reasonable, "false" otherwise):

1. a function that returns the percentage of words that start with capital letters;
2. a function that returns the presence of the phrase "you received a new message";
3. a function that performs a Google search on each word and returns whether the input contains the name of a politician.

Quiz 1 - Problem 3

Which ones of the following might be other reasonable features to use for our problem? (enter "true" if it is reasonable, "false" otherwise):

1. **a function that returns the percentage of words that start with capital letters;**
2. a function that returns the presence of the phrase "you received a new message";
3. a function that performs a Google search on each word and returns whether the input contains the name of a politician.

Yes! It can help us identify pronouns in English!

Quiz 1 - Problem 3

Which ones of the following might be other reasonable features to use for our problem? (enter "true" if it is reasonable, "false" otherwise):

1. a function that returns the percentage of words that start with capital letters;
2. **a function that returns the presence of the phrase "you received a new message";**
3. a function that performs a Google search on each word and returns whether the input contains the name of a politician.

Yes! You can argue that this can be an informative feature for indicating a message is in the **Social** or **Forum** category.

Quiz 1 - Problem 3

Which ones of the following might be other reasonable features to use for our problem? (enter "true" if it is reasonable, "false" otherwise):

1. a function that returns the percentage of words that start with capital letters;
2. a function that returns the presence of the phrase "you received a new message";
3. **a function that performs a Google search on each word and returns whether the input contains the name of a politician.**

Yes! You can also argue this can be an informative feature for **Promotions** (call to action) category. However, you can also argue that rely on an external system might not be the best idea.

Quiz 1 - Problem 4

Suppose our MLR model returns the following scores for our input:

$$\begin{aligned}
\text{score}_{\text{LMR}}(\mathbf{x}, \text{"Primary"}; \theta) &= 3.0 \\
\text{score}_{\text{LMR}}(\mathbf{x}, \text{"Social"}; \theta) &= 1.0 \\
\text{score}_{\text{LMR}}(\mathbf{x}, \text{"Promotions"}; \theta) &= 0.5 \\
\text{score}_{\text{LMR}}(\mathbf{x}, \text{"Updates"}; \theta) &= 0.1 \\
\text{score}_{\text{LMR}}(\mathbf{x}, \text{"Forums"}; \theta) &= 0.1
\end{aligned}$$

Calculate the value of the partition function $Z(\mathbf{x}, \theta)$, please round your final answer to 1 decimal place.

Quiz 1 - Problem 4

Suppose our MLR model returns the following scores for our input:

$$\begin{aligned}\text{score}_{\text{LMR}}(\mathbf{x}, \text{"Primary"}; \theta) &= 3.0\\ \text{score}_{\text{LMR}}(\mathbf{x}, \text{"Social"}; \theta) &= 1.0\\ \text{score}_{\text{LMR}}(\mathbf{x}, \text{"Promotions"}; \theta) &= 0.5\\ \text{score}_{\text{LMR}}(\mathbf{x}, \text{"Updates"}; \theta) &= 0.1\\ \text{score}_{\text{LMR}}(\mathbf{x}, \text{"Forums"}; \theta) &= 0.1\end{aligned}$$

Calculate the value of the partition function $Z(\mathbf{x}, \theta)$, please round your final answer to 1 decimal place.

$$\begin{aligned}Z(x;\theta) &= \sum_{l' \in L} \exp(score_{MLR}(x, l'; \theta))\\ &= \exp(3.0) + \exp(1.0) + \exp(0.5) + \exp(0.1) + \exp(0.1)\\ &= 26.7\ldots\end{aligned}$$

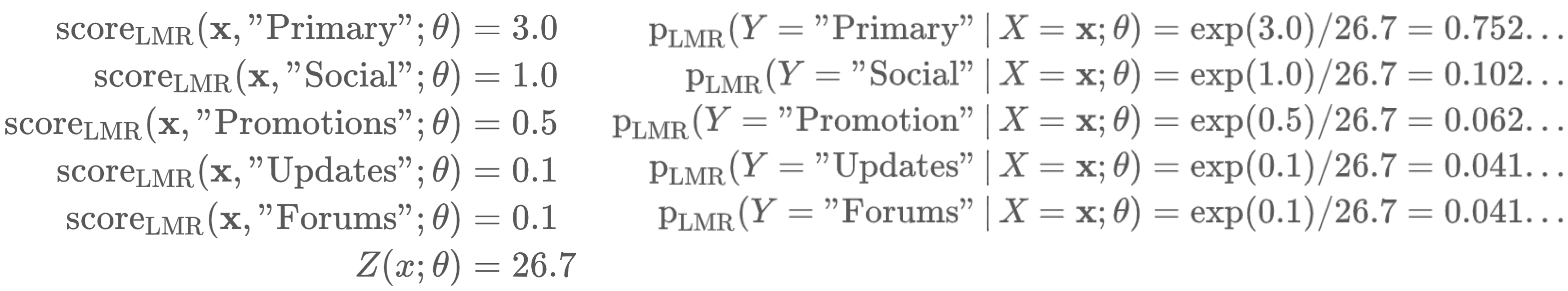
Quiz 1 - Problem 5

Calculate probability for each of the labels. Use the value you written above as the value for the partition function, and round your final answer to 3 decimal places.

$$\begin{aligned}
\text{score}_{\text{LMR}}(\mathbf{x}, ”\text{Primary}”; \theta) &= 3.0 & \text{p}_{\text{LMR}}(Y = ”\text{Primary}” \mid X = \mathbf{x}; \theta) &= \\
\text{score}_{\text{LMR}}(\mathbf{x}, ”\text{Social}”; \theta) &= 1.0 & \text{p}_{\text{LMR}}(Y = ”\text{Social}” \mid X = \mathbf{x}; \theta) &= \\
\text{score}_{\text{LMR}}(\mathbf{x}, ”\text{Promotions}”; \theta) &= 0.5 & \text{p}_{\text{LMR}}(Y = ”\text{Promotion}” \mid X = \mathbf{x}; \theta) &= \\
\text{score}_{\text{LMR}}(\mathbf{x}, ”\text{Updates}”; \theta) &= 0.1 & \text{p}_{\text{LMR}}(Y = ”\text{Updates}” \mid X = \mathbf{x}; \theta) &= \\
\text{score}_{\text{LMR}}(\mathbf{x}, ”\text{Forums}”; \theta) &= 0.1 & \text{p}_{\text{LMR}}(Y = ”\text{Forums}” \mid X = \mathbf{x}; \theta) &= \\
Z(x; \theta) &= 26.7
\end{aligned}$$

Quiz 1 - Problem 5

Calculate probability for each of the labels. Use the value you written above as the value for the partition function, and round your final answer to 3 decimal places.



$$
\begin{aligned}
\text{score}_{\text{LMR}}(\mathbf{x}, \text{''Primary''}; \theta) &= 3.0 \\
\text{score}_{\text{LMR}}(\mathbf{x}, \text{''Social''}; \theta) &= 1.0 \\
\text{score}_{\text{LMR}}(\mathbf{x}, \text{''Promotions''}; \theta) &= 0.5 \\
\text{score}_{\text{LMR}}(\mathbf{x}, \text{''Updates''}; \theta) &= 0.1 \\
\text{score}_{\text{LMR}}(\mathbf{x}, \text{''Forums''}; \theta) &= 0.1 \\
Z(x; \theta) &= 26.7
\end{aligned}
$$

$$
\begin{aligned}
\text{p}_{\text{LMR}}(Y = \text{''Primary''} \mid X = \mathbf{x}; \theta) &= \exp(3.0)/26.7 = 0.752\ldots \\
\text{p}_{\text{LMR}}(Y = \text{''Social''} \mid X = \mathbf{x}; \theta) &= \exp(1.0)/26.7 = 0.102\ldots \\
\text{p}_{\text{LMR}}(Y = \text{''Promotion''} \mid X = \mathbf{x}; \theta) &= \exp(0.5)/26.7 = 0.062\ldots \\
\text{p}_{\text{LMR}}(Y = \text{''Updates''} \mid X = \mathbf{x}; \theta) &= \exp(0.1)/26.7 = 0.041\ldots \\
\text{p}_{\text{LMR}}(Y = \text{''Forums''} \mid X = \mathbf{x}; \theta) &= \exp(0.1)/26.7 = 0.041\ldots
\end{aligned}
$$

Quiz 1 - Problem 6

Suppose the label given for this input by the human expert is "**Promotion**", what is the value of the log loss? Use the value you calculated above as your intermediate result, and round your final answer to 3 decimal places.

$$\begin{aligned}
\mathrm{p}_{\mathrm{LMR}}(Y = \text{"Primary"} \mid X = \mathbf{x}; \theta) &= \exp(3.0)/26.7 = 0.752\ldots \\
\mathrm{p}_{\mathrm{LMR}}(Y = \text{"Social"} \mid X = \mathbf{x}; \theta) &= \exp(1.0)/26.7 = 0.102\ldots \\
\mathrm{p}_{\mathrm{LMR}}(Y = \text{"Promotion"} \mid X = \mathbf{x}; \theta) &= \exp(0.5)/26.7 = 0.062\ldots \\
\mathrm{p}_{\mathrm{LMR}}(Y = \text{"Updates"} \mid X = \mathbf{x}; \theta) &= \exp(0.1)/26.7 = 0.041\ldots \\
\mathrm{p}_{\mathrm{LMR}}(Y = \text{"Forums"} \mid X = \mathbf{x}; \theta) &= \exp(0.1)/26.7 = 0.041\ldots
\end{aligned}$$

$$\text{log loss} =$$

Quiz 1 - Problem 6

Suppose the label given for this input by the human expert is "**Promotion**", what is the value of the log loss? Use the value you calculated above as your intermediate result, and round your final answer to 3 decimal places.

$$\begin{aligned}
\mathrm{p}_{\mathrm{LMR}}(Y = \text{”Primary”} \mid X = \mathbf{x}; \theta) &= \exp(3.0)/26.7 = 0.752\ldots \\
\mathrm{p}_{\mathrm{LMR}}(Y = \text{”Social”} \mid X = \mathbf{x}; \theta) &= \exp(1.0)/26.7 = 0.102\ldots \\
\mathrm{p}_{\mathrm{LMR}}(Y = \text{”Promotion”} \mid X = \mathbf{x}; \theta) &= \exp(0.5)/26.7 = 0.062\ldots \\
\mathrm{p}_{\mathrm{LMR}}(Y = \text{”Updates”} \mid X = \mathbf{x}; \theta) &= \exp(0.1)/26.7 = 0.041\ldots \\
\mathrm{p}_{\mathrm{LMR}}(Y = \text{”Forums”} \mid X = \mathbf{x}; \theta) &= \exp(0.1)/26.7 = 0.041\ldots
\end{aligned}$$

$$\begin{aligned}
\text{log loss} &= -\log\left(p_{\mathrm{LMR}}(Y = \text{”Promotion”} \mid \mathbf{X} = \mathbf{x}, \theta)\right) \\
&= -\log(0.062) \\
&= 2.781
\end{aligned}$$

Probability review

- Notation
    - X: our random variable that can take on different values
    - $\mathscr{X}$: the space of all possible events (aka values our random variable X can take on)
    - x: a specific value in $\mathscr{X}$ that X can take on

- p(x) is the shorthand for p(X = x)

- Joint probability: p(X = x, Y = y)

Probability review

- Conditional probability: $p(X = x \mid Y = y)$ – very related to joint probability

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

- Therefore, $p(X = x, Y = y) = p(X = x \mid Y = y) \cdot p(Y = y)$
  and $= p(Y = y \mid X = x) \cdot p(X = x)$

Probability review

$p(X = x, Y = y) = p(X = x) \cdot p(Y = y)$ is not always true, when is this statement false?

Maximum Likelihood Principle/Estimation

- Observations: $\boldsymbol{x}$
- Set of probability distributions that are consistent w/ assumption about the data: $\mathcal{P}$
- We want to “find the best distribution of the data given the observation”

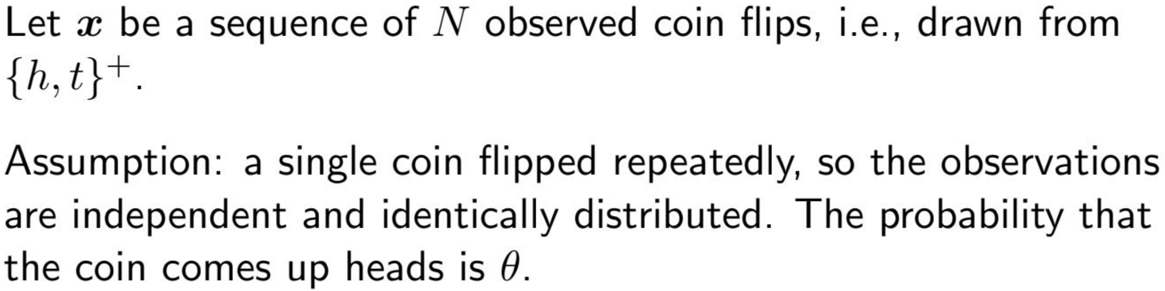
$$p_{\text{MLE}} = \operatorname*{argmax}_{p \in \mathcal{P}} p(\boldsymbol{x})$$

- In practice, we usually let $\mathcal{P}$ be a family of probabilistic models with parameter $\boldsymbol{\theta}$, we choose:

$$\boldsymbol{\theta}_{\text{MLE}} = \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{x}; \boldsymbol{\theta})$$

MLE Examples

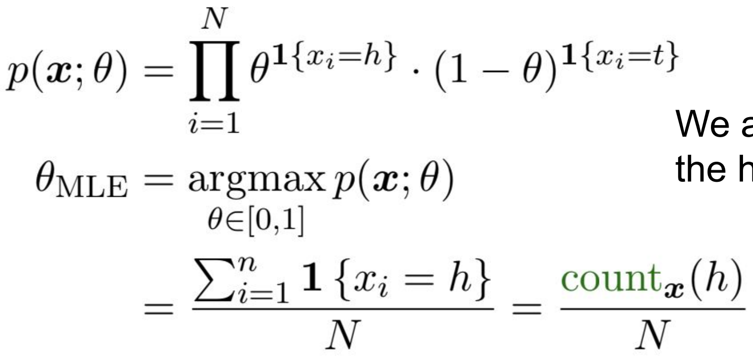
- MLE



Let $\boldsymbol{x}$ be a sequence of N observed coin flips, i.e., drawn from $\{h,t\}^+$.

Assumption: a single coin flipped repeatedly, so the observations are independent and identically distributed. The probability that the coin comes up heads is θ.

If p(coin comes up heads) = θ, then the probability of observing x given θ is



$$p(\boldsymbol{x};\theta) = \prod_{i=1}^{N} \theta^{\mathbf{1}\{x_i=h\}} \cdot (1-\theta)^{\mathbf{1}\{x_i=t\}}$$

$$\theta_{\text{MLE}} = \operatorname*{argmax}_{\theta\in[0,1]} p(\boldsymbol{x};\theta)$$

$$= \frac{\sum_{i=1}^{n} \mathbf{1}\{x_i = h\}}{N} = \frac{\text{count}_{\boldsymbol{x}}(h)}{N}$$

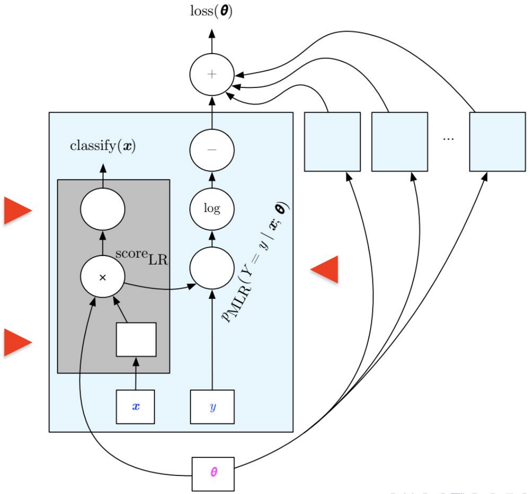
We already know how many times the heads (and the tails) came up!

Features in the multinomial setting

General template:

$$f_{\ell,\phi}(\boldsymbol{x}, y) = \phi(\boldsymbol{x}) \cdot \mathbf{1}\{y = \ell\}$$

Slide 57

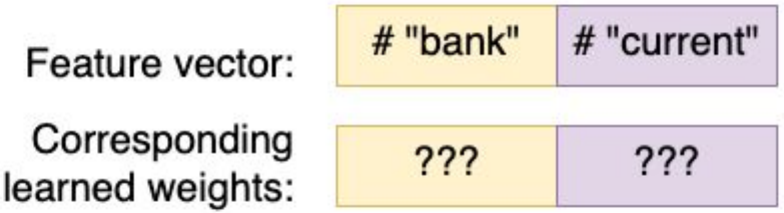

loss($\boldsymbol{\theta}$)
+
classify($\boldsymbol{x}$)
−
log
score$_{\text{LR}}$
×
$p_{\text{MLR}}(Y = y \mid \boldsymbol{x}; \boldsymbol{\theta})$
...
$\boldsymbol{x}$
y
$\boldsymbol{\theta}$

Does f depend on y?

Slide 72

Features in the multinomial setting

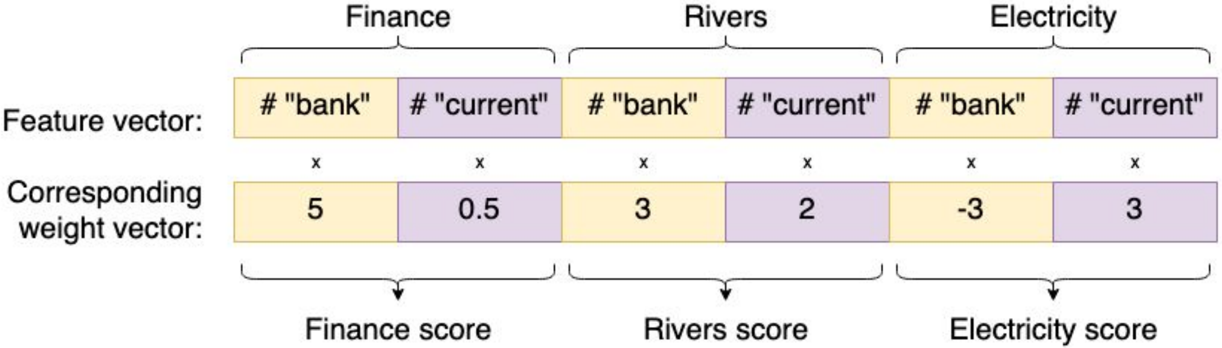
- Text classification problem: **“Categorize a document as being about exactly one of the following three topics: finance, rivers, or electricity.”**
- We think the following two words might be informative for our model: **bank** and **current**



Feature vector:	# "bank"	# "current"
Corresponding learned weights:	???	???

Features in the multinomial setting: Our solution

- Text classification problem: **“Categorize a document as being about exactly one of the following three topics: finance, rivers, or electricity.”**
- We think the following two words might be informative for our model: **bank** and **current**



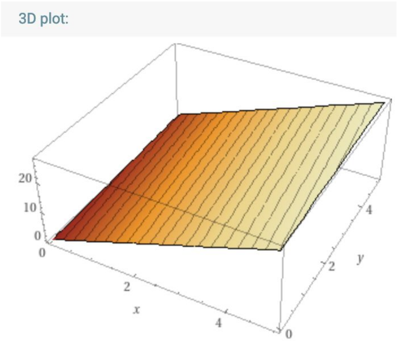
	Finance		Rivers		Electricity	
Feature vector:	# "bank"	# "current"	# "bank"	# "current"	# "bank"	# "current"
	x	x	x	x	x	x
Corresponding weight vector:	5	0.5	3	2	-3	3
	Finance score		Rivers score		Electricity score	

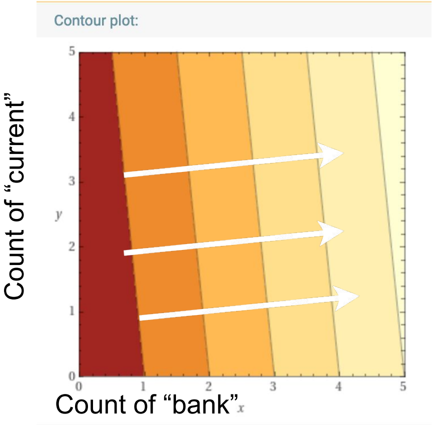
Geometric view of the problem on the previous slide

Finance:

5(bank) + 0.5(current) = score

5x + 0.5y = z

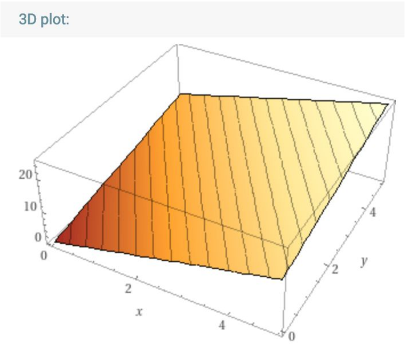


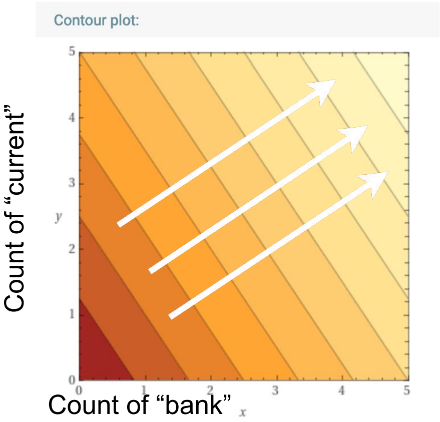


Rivers:

3(bank) + 2(current) = score

3x + 2y = z

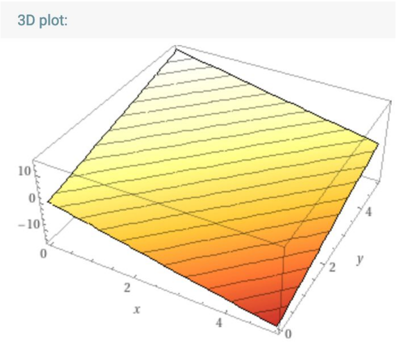


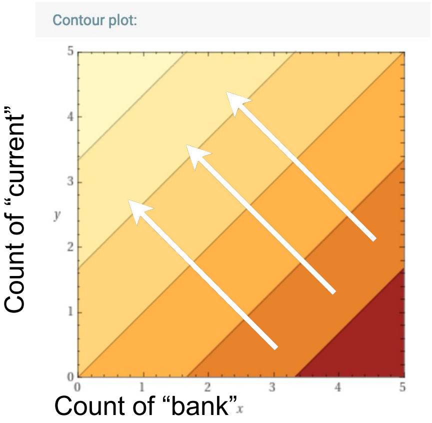


Electricity:

-3(bank) + 3(current) = score

-3x + 3y = z





Reflections from Lecture

Given what you already know about words, can you think of features that might generalize better than the ones just discussed (bag of words, presence of words, and idf)?

Reflections from Lecture

Given what you already know about words, can you think of features that might generalize better than the ones just discussed (bag of words, presence of words, and idf)?

Examples:

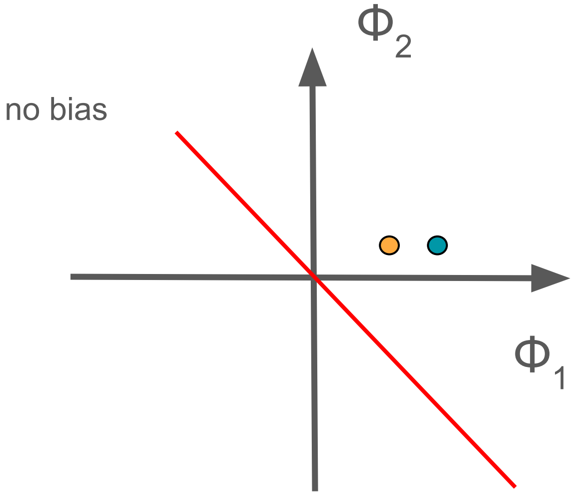
- Phrases: these give you some ordering of words, e.g. “think of”, “better than”
- Stems and morphemes: let you group up similar words, e.g. “bird” & “birds”
- Capitalization: could indicate proper nouns, e.g. # of words with capital letters
- …...

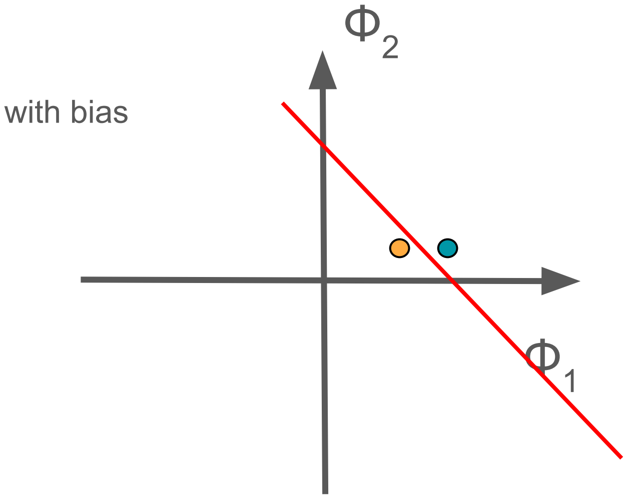
Reflections from Lecture

Recall the bias feature, $\Phi^{bias}(x) = 1$. What role does it play in the geometric interpretation of the model?

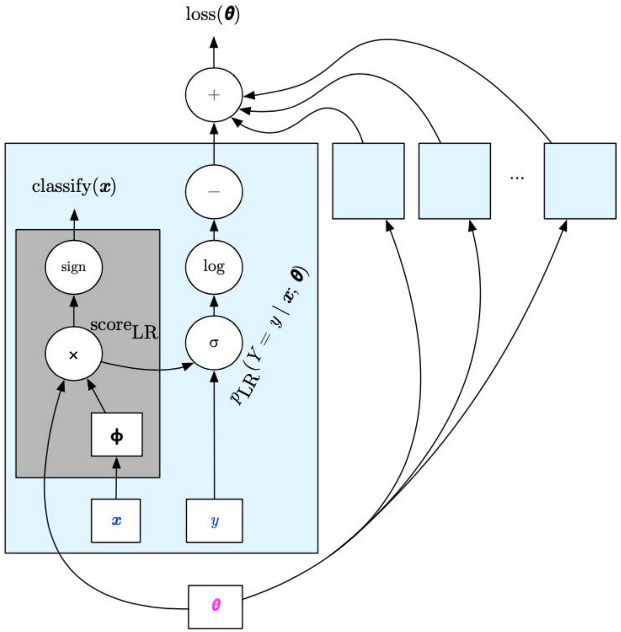
Reflections from Lecture

Recall the bias feature, $\Phi^{bias}(x) = 1$. What role does it play in the geometric interpretation of the model?





Reflections from Lecture

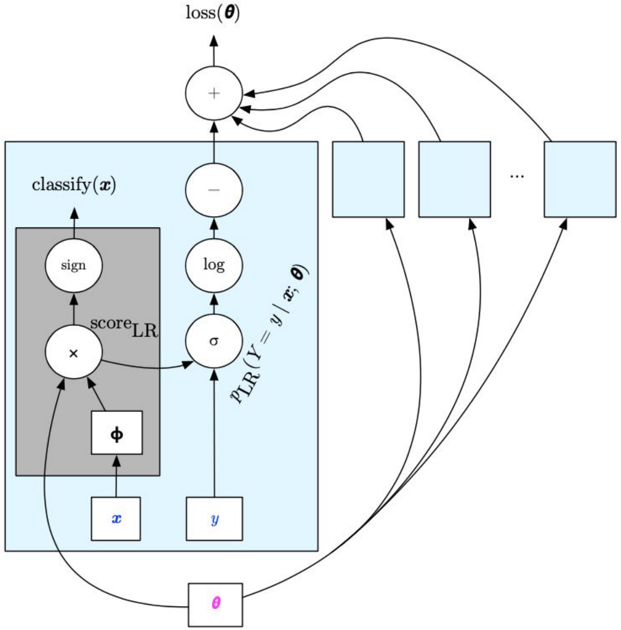

loss($\boldsymbol{\theta}$)
+
classify($\boldsymbol{x}$)
−
sign
log
score$_{\text{LR}}$
×
σ
$p_{\text{LR}}(Y = y \mid \boldsymbol{x}; \boldsymbol{\theta})$
$\boldsymbol{\phi}$
$\boldsymbol{x}$
y
...
$\boldsymbol{\theta}$

Computation graph for LR

(Lecture Slide 47)

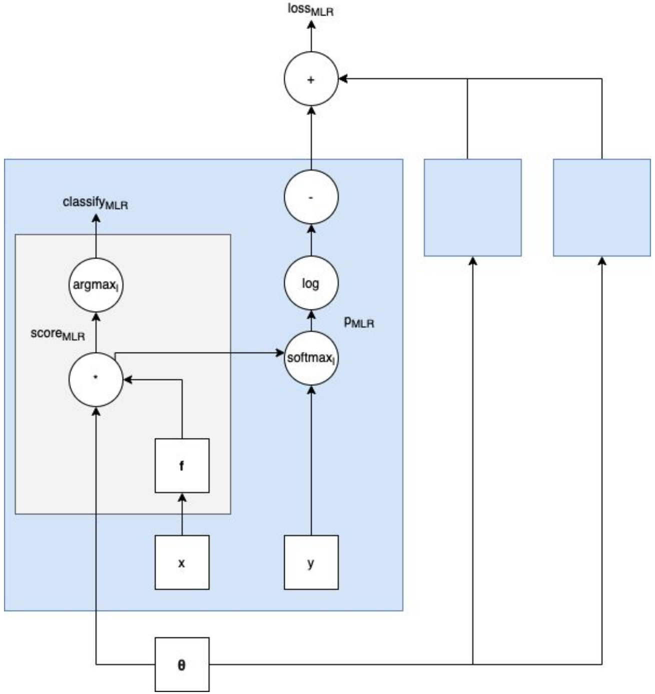
Computation graph for MLR

Reflections from Lecture


loss($\boldsymbol{\theta}$)
+
classify($\boldsymbol{x}$)
−
…
sign
log
score$_{\text{LR}}$
×
σ
$p_{\text{LR}}(Y = y \mid \boldsymbol{x}; \boldsymbol{\theta})$
ϕ
$\boldsymbol{x}$
y
$\boldsymbol{\theta}$

Computation graph for LR

(Lecture Slide 47)



Computation graph for MLR

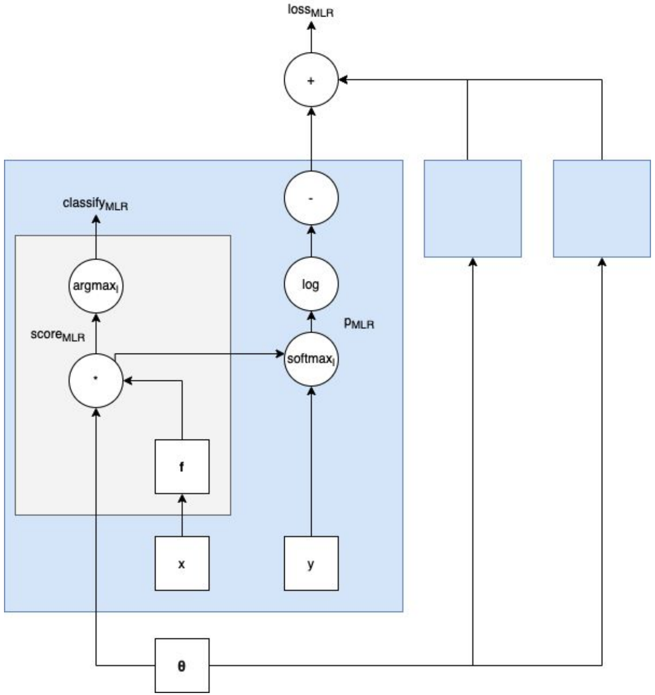
Reflections from Lecture

Computation graph of MLR

with regularization

$$\min_{\mathbf{w}} \operatorname{loss}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1$$

L1 regularization (Lecture Slide 80)



Computation graph for MLR

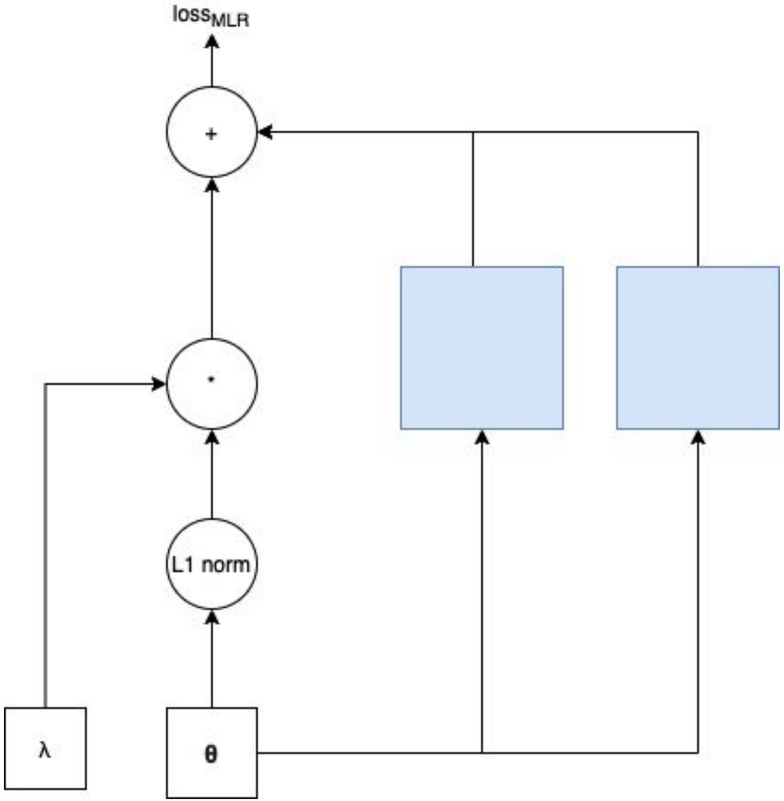
Reflections from Lecture

Computation graph of MLR

with regularization

$$\min_{\mathbf{w}} \operatorname{loss}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1$$

L1 regularization (Lecture Slide 80)



Computation graph for regularized MLR

Q & A