2 Preamble
6 Logistic Regression
23 Training: Big Picture
35 Training: Details
49 Extra: Gradient Derivation

## 25: Logistic Regression

Jerry Cain
March 8, 2024

Lecture Discussion on Ed

## Linear to Logical: Preamble

## 1. Weighted sum

Recall the linear regression model, where $\boldsymbol{X}=\left(X_{1}, X_{2}, \ldots, X_{m}\right)$ and $Y \in \mathbb{R}$ :

$$
g(\boldsymbol{X})=\theta_{0}+\sum_{j=1}^{m} \theta_{j} X_{j}
$$

How would you rewrite this expression as a single dot product?

$$
\begin{aligned}
g(\boldsymbol{X}) & =\theta_{0} X_{0}+\theta_{1} X_{1}+\theta_{2} X_{2}+\cdots+\theta_{m} X_{m} \quad \text { Define } X_{0}=1 \\
& =\theta^{T} \boldsymbol{X} \quad \text { New } \boldsymbol{X}=\left(1, X_{1}, X_{2}, \ldots, X_{m}\right)
\end{aligned}
$$

Prepending $X_{0}=1$ to each feature vector $\boldsymbol{X}$ makes matrix operators more convenient.

## 2. Sigmoid function $\sigma(z)$

- The sigmoid function:

$$
\sigma(z)=\frac{1}{1+e^{-z}}
$$

- Sigmoid squashes $z$ to a number between 0 and 1 .

- Recall definition of probability: A number between 0 and 1 that expresses a belief that something is true.
$\sigma(z)$ can represent a probability.


## 3. Conditional likelihood function

Training data ( $n$ datapoints):

- $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)$ drawn iid from a distribution $f\left(\boldsymbol{X}=\boldsymbol{x}^{(i)}, Y=y^{(i)} \mid \theta\right)=f\left(\boldsymbol{x}^{(i)}, y^{(i)} \mid \theta\right)$

$$
\begin{array}{rlr}
\theta_{M L E} & =\underset{\theta}{\arg \max } \prod_{i=1}^{n} f\left(y^{(i)} \mid x^{(i)}, \theta\right) & \begin{array}{l}
\text { conditional likelihood } \\
\text { of training data }
\end{array} \\
& =\underset{\theta}{\arg \max } \sum_{i=1}^{n} \log f\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \theta\right) & \text { Iog conditional likelihood } \\
& =\underset{\theta}{\arg \max } L L(\theta) & \begin{array}{l}
\text { - MLE here is estimator that } \\
\text { maximizes conditional likelihood } \\
\text { - Confusingly, log conditional } \\
\text { likelihood is also written as } L L(\theta)
\end{array}
\end{array}
$$

# Logistic <br> Regression 

## Prediction models so far

Linear Regression (Regression)
$X$


$$
\hat{Y}=\theta_{0}+\sum_{j=1}^{m} \theta_{j} X_{j}
$$

Naïve Bayes (Classification)

| $\boldsymbol{X}$ | $\hat{P}(\boldsymbol{X} \mid Y) \hat{P}(Y)$ |
| ---: | :--- |
| $Y$ | $\hat{P}(\boldsymbol{X}, Y)$ |
| $\hat{Y}$ | $=\underset{y=\{0,1\}}{\arg \max } P(Y \mid \boldsymbol{X})$ |
|  | $=\underset{y=\{0,1\}}{\arg \max } P(\boldsymbol{X} \mid Y) P(Y)$ |

Tractable with NB assumption, but...
! Realistically, $X_{j}$ features are not always conditionally independent
Actually models $P(\boldsymbol{X}, Y)$, not $P(Y \mid \boldsymbol{X})$ ?

## Logistic Regression

$$
\boldsymbol{X} \quad \theta_{0}+\sum_{j=1}^{m} \theta_{j} X_{j} \quad z \quad \begin{gathered}
\text { sigmoid function } \\
\sigma(z)=\frac{1}{1+e^{-z}}
\end{gathered} \quad \Rightarrow P(Y=1 \mid \boldsymbol{X})
$$

Logistic Regression Model:

$$
P(Y=1 \mid \boldsymbol{X}=\boldsymbol{x})=\sigma\left(\theta_{0}+\sum_{j=1}^{m} \theta_{j} x_{j}\right)
$$

Predict $\hat{Y}$ as the more likely $Y$

$$
\hat{Y}=\arg \max P(Y \mid \boldsymbol{X})
$$ given our observation $\boldsymbol{X}=\boldsymbol{x}$ :

$$
y=\{0,1\}
$$

- Since $Y \in\{0,1\}$,

$$
P(Y=0 \mid \boldsymbol{X}=\boldsymbol{x})=1-\sigma\left(\theta_{0}+\sum_{j=1}^{m} \theta_{j} x_{j}\right)
$$

- Sigmoid function also known as logit function


## Logistic Regression



$$
P(Y=1 \mid \boldsymbol{X}=x)=\sigma\left(\theta_{0}+\sum_{j=1}^{m} \theta_{j} x_{j}\right)
$$

## Logistic Regression: Key Metaphor


$\theta$ parameter

## Logistic Regression: Key Metaphor



## Logistic Regression: Key Metaphor


[0,1,1]

## Components of Logistic Regression



## Components of Logistic Regression



## Components of Logistic Regression



## Components of Logistic Regression



## Different predictions for different inputs


[0,1,1]

## Different predictions for different inputs


[0,0,1]

## Parameters affect prediction



## Parameters affect prediction



## Parameters affect prediction

$$
\begin{gathered}
P(Y=1 \mid \boldsymbol{X}=\boldsymbol{x})=\sigma\left(\theta_{0}+\sum_{j=1}^{m} \theta_{j} x_{j}\right) \\
P(Y=1 \mid \boldsymbol{X}=\boldsymbol{x})=\sigma\left(\sum_{j=0}^{m} \theta_{j} x_{j}\right)=\sigma\left(\theta^{T} \boldsymbol{x}\right) \quad \text { where } x_{0}=1
\end{gathered}
$$

## Logistic regression classifier

$$
\begin{aligned}
& \hat{Y}=\underset{y=\{0,1\}}{\arg \max } P(Y \mid \boldsymbol{X}) \\
& P(Y=1 \mid \boldsymbol{X}=\boldsymbol{x})=\sigma\left(\sum_{j=0}^{m} \theta_{j} x_{j}\right)=\sigma\left(\theta^{T} \boldsymbol{x}\right)
\end{aligned}
$$

## Training

Estimate parameters
from training data

$$
\theta=\left(\theta_{0}, \theta_{1}, \theta_{2}, \ldots, \theta_{m}\right)
$$

Testing
Given an observation $\boldsymbol{X}=\left(X_{1}, X_{2}, \ldots, X_{m}\right)$, predict

$$
\hat{Y}=\arg \max P(Y \mid \boldsymbol{X})
$$

$$
y=\{0,1\}
$$

## Training: The big picture

## Logistic regression classifier

$$
\begin{aligned}
& \hat{Y}=\underset{y=\{0,1\}}{\arg \max } P(Y \mid \boldsymbol{X}) \\
& P(Y=1 \mid \boldsymbol{X}=\boldsymbol{x})=\sigma\left(\sum_{j=0}^{m} \theta_{j} x_{j}\right)=\sigma\left(\theta^{T} \boldsymbol{x}\right)
\end{aligned}
$$

## Training

Estimate parameters
from training data

$$
\theta=\left(\theta_{0}, \theta_{1}, \theta_{2}, \ldots, \theta_{m}\right)
$$

Choose $\theta$ that optimizes some objective:

1. Determine objective function
2. Find gradient with respect to each $\theta$
3. Solve analytically by setting to 0 , or solve computationally with gradient ascent

We are modeling $P(Y \mid X)$ directly, so we maximize the conditional likelihood of training data.

## Estimating $\theta$

1. Determine objective function

$$
\theta_{M L E}=\underset{\theta}{\arg \max } \prod_{i=1}^{n} f\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \theta\right)
$$

2. Gradient wrt $\theta_{j}$, for $j=0,1, \ldots, m$
3. Solve

- No analytical derivation of $\theta_{M L E} \ldots$
- ...but can still determine $\theta_{\text {MLE }}$ via gradient ascent!

$$
\begin{aligned}
& \text { initialize } x \\
& \text { repeat many times: } \\
& \text { compute gradient } \\
& x+=\eta \text { * gradient }
\end{aligned}
$$

## 1. Determine objective function

$\theta_{M L E}=$| $\left.\underset{\theta}{\arg \max } \prod_{i=1}^{n} f\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \theta\right)=\begin{array}{r}P(Y=1 \mid \boldsymbol{X}=\boldsymbol{x})=\sigma\left(\sum_{j=0}^{m} \theta_{j} x_{j}\right) \\ \arg \max \\ \theta\end{array}\right)$ |
| :---: |

First: Interpret conditional likelihood with Logistic Regression

Second: Write a differentiable expression for log conditional likelihood

## 1. Determine objective function (interpret)

$\theta_{M L E}=\underset{\theta}{\arg \max } \prod_{i=1}^{n} f\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \theta\right)=\underset{\theta}{\arg \max } L L(\theta) \quad \begin{aligned} P(Y=1 \mid \boldsymbol{X}=\boldsymbol{x}) & =\sigma\left(\sum_{j=0}^{m} \theta_{j} x_{j}\right) \\ & =\sigma\left(\theta^{T} \boldsymbol{x}\right)\end{aligned}$
Suppose you have $n=2$ training datapoints: $\quad\left(\boldsymbol{x}^{(1)}, 1\right),\left(\boldsymbol{x}^{(2)}, 0\right)$
Consider the following expressions for a given $\theta$ :
A. $\sigma\left(\theta^{T} \boldsymbol{x}^{(1)}\right) \sigma\left(\theta^{T} \boldsymbol{x}^{(2)}\right)$
B. $\left(1-\sigma\left(\theta^{T} \boldsymbol{x}^{(1)}\right)\right) \sigma\left(\theta^{T} \boldsymbol{x}^{(2)}\right)$
C. $\sigma\left(\theta^{T} \boldsymbol{x}^{(1)}\right)\left(1-\sigma\left(\theta^{T} \boldsymbol{x}^{(2)}\right)\right)$
D. $\left(1-\sigma\left(\theta^{T} \boldsymbol{x}^{(1)}\right)\right)\left(1-\sigma\left(\theta^{T} \boldsymbol{x}^{(2)}\right)\right)$

1. Interpret the above expressions as probabilities.
2. If we let $\theta=\theta_{M L E}$, which probability should be the highest?

## 1. Determine objective function (write)

$\theta_{M L E}=\underset{\theta}{\arg \max } \prod_{i=1}^{n} f\left(y^{(i)} \mid x^{(i)}, \theta\right)=\underset{\theta}{\arg \max } L L(\theta)$

$$
\begin{aligned}
P(Y=1 \mid \boldsymbol{X}=\boldsymbol{x}) & =\sigma\left(\sum_{j=0}^{m} \theta_{j} x_{j}\right) \\
& =\sigma\left(\theta^{T} \boldsymbol{x}\right)
\end{aligned}
$$

1. What is a differentiable expression for $P(Y=y \mid \boldsymbol{X}=\boldsymbol{x})$ ?

$$
P(Y=y \mid \boldsymbol{X}=\boldsymbol{x})= \begin{cases}\sigma\left(\theta^{T} \boldsymbol{x}\right) & \text { if } y=1 \\ 1-\sigma\left(\theta^{T} \boldsymbol{x}\right) & \text { if } y=0\end{cases}
$$

## Recall

Bernoulli
MLE!
2. What is a differentiable expression for $L L(\theta)$, log conditional likelihood?

$$
L L(\theta)=\log \prod_{i=1}^{n} f\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \theta\right)
$$

## 1. Determine objective function (write)

$\theta_{M L E}=\underset{\theta}{\arg \max } \prod_{i=1}^{n} f\left(y^{(i)} \mid x^{(i)}, \theta\right)=\underset{\theta}{\arg \max } L L(\theta)$

$$
\begin{aligned}
P(Y=1 \mid \boldsymbol{X}=\boldsymbol{x}) & =\sigma\left(\sum_{j=0}^{m} \theta_{j} x_{j}\right) \\
& =\sigma\left(\theta^{T} \boldsymbol{x}\right)
\end{aligned}
$$

1. What is a differentiable expression for $P(Y=y \mid \boldsymbol{X}=\boldsymbol{x})$ ?

$$
P(Y=y \mid \boldsymbol{X}=\boldsymbol{x})=\left(\sigma\left(\theta^{T} \boldsymbol{x}\right)\right)^{y}\left(1-\sigma\left(\theta^{T} \boldsymbol{x}\right)\right)^{1-y}
$$

2. What is a differentiable expression for $L L(\theta)$, log conditional likelihood?

$$
L L(\theta)=\sum_{i=1}^{n} y^{(i)} \log \sigma\left(\theta^{T} x^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-\sigma\left(\theta^{T} x^{(i)}\right)\right)
$$

## 2. Find gradient with respect to $\theta$

Optimization problem:

$$
\begin{gathered}
\theta_{M L E}=\underset{\theta}{\arg \max } \prod_{i=1}^{n} f\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \theta\right)=\underset{\theta}{\arg \max } L L(\theta) \\
L L(\theta)=\sum_{i=1}^{n} y^{(i)} \log \sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right)
\end{gathered}
$$

Gradient wrt $\theta_{j}$, for $j=0,1, \ldots, m$ :

$$
\frac{\partial L L(\theta)}{\partial \theta_{j}}=\sum_{i=1}^{n}\left[y^{(i)}-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right] x_{j}^{(i)}
$$

** presented here without proof, though it is ***
a generalization of the gradient of a in the
$\mathrm{Y}=a X+b+Z$ derivation from last lecture

How do we interpret the gradient contribution of the $i^{\text {th }}$ training datapoint?

## 2. Find gradient with respect to $\theta$

Optimization problem:

$$
\begin{gathered}
\theta_{M L E}=\underset{\theta}{\arg \max } \prod_{i=1}^{n} f\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \theta\right)=\underset{\theta}{\arg \max } L L(\theta) \\
L L(\theta)=\sum_{i=1}^{n} y^{(i)} \log \sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right)
\end{gathered}
$$

Gradient wrt $\theta_{j}$, for $j=0,1, \ldots, m$ :

$$
\frac{\partial L L(\theta)}{\partial \theta_{j}}=\sum_{i=1}^{n}\left[y^{(i)}-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right] x_{j}^{(i)}
$$

## 2. Find gradient with respect to $\theta$

Optimization problem:

$$
\begin{gathered}
\theta_{M L E}=\underset{\theta}{\arg \max } \prod_{i=1}^{n} f\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \theta\right)=\underset{\theta}{\arg \max } L L(\theta) \\
L L(\theta)=\sum_{i=1}^{n} y^{(i)} \log \sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right)
\end{gathered}
$$

Gradient wrt $\theta_{j}$, for $j=0,1, \ldots, m$ :

$$
\begin{aligned}
\frac{\partial L L(\theta)}{\partial \theta_{j}}= & \sum_{i=1}^{n}\left[y^{(i)}-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right] x_{j}^{(i)} \\
& 1 \text { or } 0 \quad P\left(Y=1 \mid X=x^{(i)}\right)
\end{aligned}
$$

## 2. Find gradient with respect to $\theta$

Optimization problem:

$$
\begin{gathered}
\theta_{M L E}=\underset{\theta}{\arg \max } \prod_{i=1}^{n} f\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \theta\right)=\underset{\theta}{\arg \max } L L(\theta) \\
L L(\theta)=\sum_{i=1}^{n} y^{(i)} \log \sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right)
\end{gathered}
$$

Gradient wrt $\theta_{j}$, for $j=0,1, \ldots, m$ :

$$
\frac{\partial L L(\theta)}{\partial \theta_{j}}=\sum_{i=1}^{n} \underbrace{\left[y^{(i)}-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right]} x_{j}^{(i)}
$$

Suppose $y^{(i)}=1$ (the true class label for the $i^{\text {th }}$ datapoint):

- If $\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right) \geq 0.5$, correct
- If $\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)<0.5$, incorrect $\rightarrow$ change $\theta_{j}$ more


## 3. Solve

1. Optimization problem:

$$
\begin{gathered}
\theta_{M L E}=\underset{\theta}{\arg \max } \prod_{i=1}^{n} f\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \theta\right)=\underset{\theta}{\arg \max } L L(\theta) \\
L L(\theta)=\sum_{i=1}^{n} y^{(i)} \log \sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right)
\end{gathered}
$$

2. Gradient wrt $\theta_{j}$, for $j=0,1, \ldots, m: \quad \frac{\partial L L(\theta)}{\partial \theta_{j}}=\sum_{i=1}^{n}\left[y^{(i)}-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right] x_{j}^{(i)}$
3. Solve using gradient ascent!

## Training: The details

## Training: Gradient ascent step

$$
\frac{\partial L L(\theta)}{\partial \theta_{j}}=\sum_{i=1}^{n}\left[y^{(i)}-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right] x_{j}^{(i)} \quad \text { for } j=0,1, \ldots, m
$$

repeat until convergence:
for all thetas:

$$
\begin{aligned}
\theta_{j}^{\text {new }} & =\theta_{j}^{\text {old }}+\eta \cdot \frac{\partial L L\left(\theta^{\text {old }}\right)}{\partial \theta_{j}^{\text {old }}} \\
& =\theta_{j}^{\text {old }}+\eta \cdot \sum_{i=1}^{n}\left[y^{(i)}-\sigma\left(\theta^{\text {old }^{T}} \boldsymbol{x}^{(i)}\right)\right] x_{j}^{(i)} \quad \begin{array}{c}
\text { What does } \\
\text { this look like } \\
\text { in code? }
\end{array}
\end{aligned}
$$

## Training: Gradient Ascent

$$
\text { for } j=0,1, \ldots, m \text { : }
$$

$$
\begin{aligned}
& \text { Gradient } \theta_{j}^{\text {new }}=\theta_{j}^{\text {old }}+\eta \cdot \sum_{i=1}^{n}\left[y^{(i)}-\sigma\left(\theta^{\text {old }^{T}} \boldsymbol{x}^{(i)}\right)\right] x_{j}^{(i)} \\
& \text { Ascent Step }
\end{aligned}
$$

initialize $\theta_{j}=0$ for $0 \leq j \leq m$ repeat until convergence:

```
gradient[j] = 0 for 0 \leq j \leqm
// TODO: your code here
// compute all gradient[j]'s
// based on n training examples
```

$\theta_{j}+=\eta * \operatorname{gradient}[j]$ for all $0 \leq j \leq m$

## Training: Gradient Ascent <br> 

initialize $\theta_{j}=0$ for $0 \leq j \leq m$ repeat until convergence:

$$
\begin{aligned}
& \text { gradient }[j]=0 \text { for } 0 \leq j \leq m \\
& \text { for each training example }(x, y) \text { : } \\
& \text { for each } 0 \leq j \leq m \text { : } \\
& \text { // update gradient }[j] \text { for } \\
& \text { // current }(x, y) \text { example } \\
& \theta_{j}+=\eta \text { * gradient }[j] \text { for all } 0 \leq j \leq m
\end{aligned}
$$


initialize $\theta_{j}=0$ for $0 \leq j \leq m$ repeat until convergence :

$$
\operatorname{gradient}[j]=0 \text { for } 0 \leq j \leq m
$$

$$
\text { for each training example }(x, y) \text { : }
$$

$$
\text { for each } 0 \leq j \leq m \text { : }
$$

$$
\text { gradient }[\mathrm{j}]+=\left[y-\frac{1}{1+e^{-\theta^{T} x}}\right] x_{j}
$$

Some important details...

$$
\theta_{j}+=\eta * \operatorname{gradient}[\mathrm{j}] \text { for all } 0 \leq \mathrm{j} \leq m
$$

## Training: Gradient Ascent

$$
\begin{array}{r}
\text { Gradient } \\
\text { scent Step }
\end{array} \theta_{j}^{\text {new }}=\theta_{j}^{\text {old }}+\eta \cdot \sum_{i=1}^{n}\left[y^{(i)}-\sigma\left(\theta^{\text {old }^{T}} x^{(i)}\right)\right] x_{j}^{(i)}
$$

initialize $\theta_{j}=0$ for $0 \leq \mathrm{j} \leq \mathrm{m}$ repeat until convergence:

$$
\operatorname{gradient}[j]=0 \text { for } 0 \leq j \leq m
$$

for each training example $(x, y)$ :
for each $0 \leq \mathrm{j} \leq \mathrm{m}$ :

$$
\text { gradient }[\mathrm{j}]+=\left[y-\frac{1}{1+e^{-\theta^{T} x}}\right] x_{j}
$$

$$
\theta_{j}+=\eta * \text { gradient }[j] \text { for all } 0 \leq j \leq m
$$

## Training: Gradient Ascent

initialize $\theta_{j}=0$ for $0 \leq j \leq m$ repeat until convergence:

$$
\text { gradient[j] }=0 \text { for } 0 \leq j \leq m
$$

for each training example $(x, y)$ :
for each $0 \leq j \leq m$ :
gradient $[j]+=\left[y-\frac{1}{1+e^{-\theta^{T} x}}\right] x_{j}$
$\theta_{j}+=\cap$ gradient[j] for all $0 \leq j \leq m$

- Finish computing gradient with $\theta^{\text {old }}$ prior to any $\theta$ update
- Learning rate $\eta$ is a constant you set before training


## Training: Gradient Ascent

initialize $\theta_{j}=0$ for $0 \leq j \leq m$ repeat until convergence:

$$
\text { gradient [j] }=0 \text { for } 0 \leq j \leq m
$$

for each training example $(x, y)$ :
for each $0 \leq j \leq m$ :

$$
\text { gradient }[\mathrm{j}]+=\left[y-\frac{1}{1+e^{-\theta^{T} x}}\right] x_{j}
$$

$\theta_{j}+=\eta * \operatorname{gradient[j]}$ for all $0 \leq j \leq m$

- Finish computing gradient with $\theta^{\text {old }}$ prior to any $\theta$ update
- Learning rate $\eta$ is a constant you set before training
- $x_{j}$ is the $j^{\text {th }}$ feature of input $\boldsymbol{x}=\left(x_{1}, \ldots, x_{m}\right)$


## Training: Gradient Ascent

initialize $\theta_{j}=0$ for $0 \leq j \leq m$ repeat until convergence:

$$
\text { gradient[j] }=0 \text { for } 0 \leq j \leq m
$$

for each training example $(x, y)$ :
for each $0 \leq j \leq m:$

$$
\operatorname{gradient}[\mathrm{j}]+=\left[y-\frac{1}{\left.1+e^{-\theta^{T} x}\right]} x_{j}\right.
$$

$\theta_{j}+=\eta * \operatorname{gradient}[j]$ for all $0 \leq j \leq m$

- Finish computing gradient with $\theta^{\text {old }}$ prior to any $\theta$ update
- Learning rate $\eta$ is a constant you set before training
- $x_{j}$ is the $j^{\text {th }}$ feature of input $\boldsymbol{x}=\left(x_{1}, \ldots, x_{m}\right)$
- Insert $x_{0}=1$ before training


## Training: Gradient Ascent

| Gradient |
| :---: |
| Ascent Step |
| $\theta_{j}^{\text {new }}$ |$=\theta_{j}^{\text {old }}+\eta \cdot \sum_{i=1}^{n}\left[y^{(i)}-\sigma\left(\theta^{\text {old }^{T}} \boldsymbol{x}^{(i)}\right)\right] x_{j}^{(i)}$

initialize $\theta_{j}=0$ for $0 \leq j \leq m$ repeat until convergence:

$$
\operatorname{gradient}[j]=0 \text { for } 0 \leq j \leq m
$$

for each training example $(x, y)$ :
for each $0 \leq j \leq m:$

$$
\text { gradient }[\mathrm{j}]+=\left[y-\frac{1}{1+e^{-\theta^{T} x}}\right] x_{j}
$$

$\theta_{j}+=\eta * \operatorname{gradient}[j]$ for all $0 \leq j \leq m$

- Finish computing gradient with $\theta^{\text {old }}$ prior to any $\theta$ update
- Learning rate $\eta$ is a constant you set before training
- $x_{j}$ is the $j^{\text {th }}$ feature of input $\boldsymbol{x}=\left(x_{1}, \ldots, x_{m}\right)$
- Insert $x_{0}=1$ before training


## Naïve Bayes

## Logistic Regression



Compare/contrast:

1. What distributions are we modeling?
2. After learning our parameters, could we randomly generate a new datapoint $(\boldsymbol{x}, y)$ ?
3. Could we model a continuous $X_{j}$ feature (e.g., $X_{j} \sim$ Normal, or $X_{j} \sim$ Unknown)?
4. Could we model a non-binary discrete $X_{j}$ (e.g., $X_{j} \in\{1,2, \ldots, 6\}$ )?

## Tradeoffs:

1. Modeling goal
2. Generative or discriminative?
3. Continuous input features

Naïve Bayes
$P(\boldsymbol{X}, Y)$
Generative: could use joint distribution to generate new points (! but you might not need this extra effort)

4 Needs parametric form (e.g., Gaussian) or discretized buckets (for multinomial features)
4. Discrete input features

## Logistic Regression

$$
P(Y \mid \boldsymbol{X})
$$

Discriminative: just tries to discriminate $y=0$ vs $y=1$ ( $\boldsymbol{X}$ cannot generate new points b/c no $P(\boldsymbol{X}, Y))$
$\nabla$ Yes, easily
! Multi-valued discrete data hard (e.g., if $X_{i} \in\{A, B, C\}$, not necessarily good to encode as
$\{1,2,3\}$

## Linearly separable data

Logistic Regression is trying to find the line that separates data instances where $y=1$ from those where $y=0$ :


- We call such data (or functions generating that data) linearly separable.
- Naïve Bayes is linear too, because there is
 one parameter for each feature (and no parameters that involve multiple features).


## Data is not always linearly separable



- Not possible to draw a line that successfully separates all the $y=1$ points (green) from the $y=0$ points (red)
- Despite this, Logistic Regression and Naive Bayes still often work well in practice


# Extra: Gradient Derivation 

## Background: Calculus

## Calculus refresher <br> \#1: <br> Derivative(sum) = <br> sum(derivative)

## Calculus refresher

\#2:
Chain rule

$$
\frac{\partial}{\partial x} \sum_{i=1}^{n} f_{i}(x)=\sum_{i=1}^{n} \frac{\partial f_{i}(x)}{\partial x}
$$

$$
\begin{aligned}
\frac{\partial f(x)}{\partial x}=\frac{\partial f(z)}{\partial z} \frac{\partial z}{\partial x} & \text { Calculus Chain Rule } \\
f(x)=f(z(x)) & \begin{array}{l}
\text { aka decomposition } \\
\text { of composed functions }
\end{array}
\end{aligned}
$$

## Our goal

Find: $\frac{\partial L L(\theta)}{\partial \theta_{j}}$ where

$$
L L(\theta)=\sum_{i=1}^{n} y^{(i)} \log \sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right) \quad \begin{aligned}
& \text { log conditional } \\
& \text { likelihood }
\end{aligned}
$$

Two "pre-processing" steps to prepare for chain rule

1. Rewrite $L L(\theta)$ with $\hat{y}$
2. Compute gradient of $\hat{y}$

## 1. Rewriting $L L(\theta)$ with $\hat{y}$

Find: $\frac{\partial L L(\theta)}{\partial \theta_{j}}$ where

$$
\begin{aligned}
& L L(\theta)=\sum_{i=1}^{n} y^{(i)} \log \sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right) \quad \begin{array}{l}
\text { log conditional } \\
\text { likelihood }
\end{array} \\
& L L(\theta)=\sum_{i=1}^{n} y^{(i)} \log \hat{y}^{(i)}+\left(1-y^{(i)}\right) \log \left(1-\hat{y}^{(i)}\right) \quad \text { Let } \hat{y}^{(i)}=\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)
\end{aligned}
$$

## 2. Compute gradient of $\hat{\boldsymbol{y}}=\sigma\left(\theta^{T} \boldsymbol{x}\right)$

Aside: Sigmoid has a beautiful derivative!

Sigmoid function:

$$
\sigma(z)=\frac{1}{1+e^{-z}}
$$

Derivative:

$$
\frac{d}{d z} \sigma(z)=\sigma(z)[1-\sigma(z)]
$$

## 2. Compute gradient of $\hat{\boldsymbol{y}}=\sigma\left(\theta^{T} \boldsymbol{x}\right)$

Sigmoid function:

$$
\sigma(z)=\frac{1}{1+e^{-z}}
$$

$$
\frac{d}{d z} \sigma(z)=\sigma(z)[1-\sigma(z)]
$$

What is $\frac{\partial}{\partial \theta_{j}} \hat{y}=\frac{\partial}{\partial \theta_{j}} \sigma\left(\theta^{T} \boldsymbol{x}\right) ?$
A. $\sigma\left(x_{j}\right)\left[1-\sigma\left(x_{j}\right)\right] x_{j}$
B. $\sigma\left(\theta^{T} \boldsymbol{x}\right)\left[1-\sigma\left(\theta^{T} \boldsymbol{x}\right)\right] \boldsymbol{x}$
C. $\sigma\left(\theta^{T} \boldsymbol{x}\right)\left[1-\sigma\left(\theta^{T} \boldsymbol{x}\right)\right] x_{j}$
D. $\sigma\left(\theta^{T} \boldsymbol{x}\right) x_{j}\left[1-\sigma\left(\theta^{T} \boldsymbol{x}\right) x_{j}\right]$
E. None/other

## 2. Compute gradient of $\hat{\boldsymbol{y}}=\sigma\left(\theta^{T} \boldsymbol{x}\right)$

Sigmoid function:

$$
\sigma(z)=\frac{1}{1+e^{-z}}
$$

Derivative:

$$
\frac{d}{d z} \sigma(z)=\sigma(z)[1-\sigma(z)]
$$

What is $\frac{\partial}{\partial \theta_{j}} \sigma\left(\theta^{T} \boldsymbol{x}\right)$ ?

$$
\text { Let } z=\theta^{T} \boldsymbol{x}=\sum_{k=0}^{m} \theta_{k} x_{k}
$$

## A. $\sigma\left(x_{j}\right)\left[1-\sigma\left(x_{j}\right)\right] x_{j}$

B. $\sigma\left(\theta^{T} x\right)\left[1-\sigma\left(\theta^{T} x\right)\right] x$

$$
\sigma\left(\theta^{T} \boldsymbol{x}\right)\left[1-\sigma\left(\theta^{T} \boldsymbol{x}\right)\right] x_{j}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \theta_{j}} \sigma\left(\theta^{T} \boldsymbol{x}\right) & =\frac{\partial}{\partial z} \sigma(z) \cdot \frac{\partial z}{\partial \theta_{j}} \quad \text { (Chain Rule) } \\
& =\sigma\left(\theta^{T} \boldsymbol{x}\right)\left[1-\sigma\left(\theta^{T} \boldsymbol{x}\right)\right] x_{j}
\end{aligned}
$$

D. $\sigma\left(\theta^{T} \boldsymbol{x}\right) x_{j}\left[1-\sigma\left(\theta^{T} \boldsymbol{x}\right) x_{j}\right]$
E. None/other

## Compute gradient of log conditional likelihood

$$
\begin{aligned}
\frac{\partial L L(\theta)}{\partial \theta_{j}} & =\sum_{i=1}^{n} \frac{\partial}{\partial \theta_{j}}\left[y^{(i)} \log \left(\hat{y}^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-\hat{y}^{(i)}\right)\right] \quad \text { Let } \hat{y}^{(i)}=\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right) \\
& =\sum_{i=1}^{n} \frac{\partial}{\partial \hat{y}^{(i)}}\left[y^{(i)} \log \left(\hat{y}^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-\hat{y}^{(i)}\right)\right] \cdot \frac{\partial \hat{y}^{(i)}}{\partial \theta_{j}} \quad \quad \text { (Chain Rule) } \\
& =\sum_{i=1}^{n}\left[y^{(i)} \frac{1}{\hat{y}^{(i)}}-\left(1-y^{(i)}\right) \frac{1}{\left.1-\hat{y}^{(i)}\right] \cdot \hat{y}^{(i)}\left(1-\hat{y}^{(i)}\right) x_{j}^{(i)}}\right. \\
& =\sum_{i=1}^{n}\left[y^{(i)}-\hat{y}^{(i)}\right] x_{j}^{(i)} \quad=\sum_{i=1}^{n}\left[y^{(i)}-\sigma\left(\theta^{T} x^{(i)}\right)\right] x_{j}^{(i)} \quad \text { (calculus) }
\end{aligned}
$$

## Compute gradient of log conditional likelihood

$$
\begin{aligned}
& \frac{\partial L L(\theta)}{\partial \theta_{j}}=\sum_{i=1}^{n} \frac{\partial}{\partial \theta_{j}}\left[y^{(i)} \log \left(\hat{y}^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-\hat{y}^{(i)}\right)\right] \quad \text { Let } \hat{y}^{(i)}=\sigma\left(\theta^{T} x^{(i)}\right) \\
& =\sum_{i=1}^{n} \frac{\partial}{\partial \hat{y}^{(i)}}\left[y^{(i)} \log \left(\hat{y}^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-\hat{y}^{(i)}\right)\right] \cdot \frac{\partial \hat{y}^{(i)}}{\partial \theta_{j}} \\
& =\sum_{i=1}^{n}\left[y^{(i)} \frac{1}{\hat{y}^{(i)}}-\left(1-y^{(i)}\right) \frac{1}{1-\hat{y}^{(i)}}\right] \cdot \hat{y}^{(i)}\left(1-\hat{y}^{(i)}\right) x_{j}^{(i)} \\
& =\sum_{i=1}^{n}\left[y^{(i)}-\hat{y}^{(i)}\right] x_{j}^{(i)} \quad=\sum_{i=1}^{n}\left[y^{(i)}-\sigma\left(\theta^{T} \boldsymbol{x}^{(i)}\right)\right] x_{j}^{(i)}
\end{aligned}
$$

