CS 383 – Machine Learning

Logistic Regression

Slides adapted from material created by E. Alpaydin
Prof. Mordohai, Prof. Greenstadt, Pattern Classification (2nd Ed.),
Pattern Recognition and Machine Learning

Objectives

- Logistic Regression

Logistic Regression

Logistic Regression

- Logistic Regression is a terrible name!
  - It's not regression at all!
  - It's classification
- But as you'll see, how we do it is extremely similar to *linear* regression

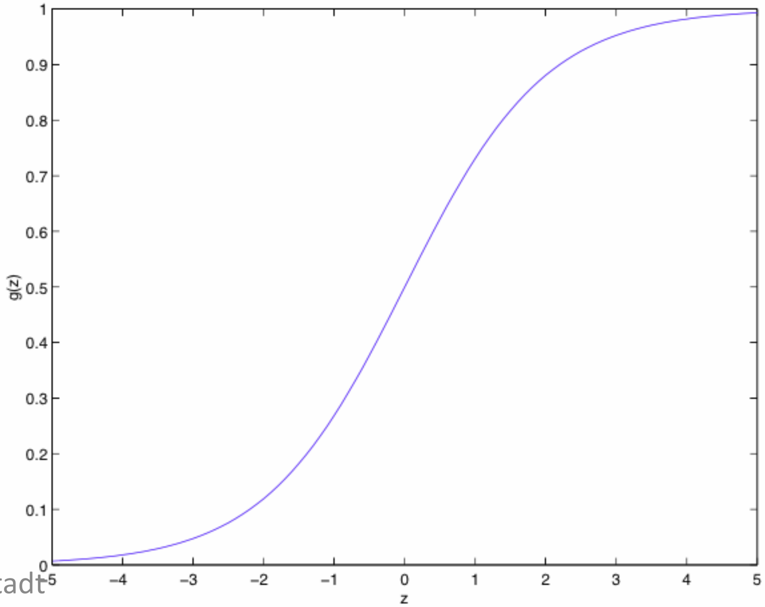
Logistic Regression

- Adapt linear regression for binary classification
$$y \in \{0,1\}$$
- Outputs a probability: $0 \leq P(y=1) \leq 1$
- Recall from *linear* regression we computed
$$g(x) = x\theta$$
- We can adapt this for use in computing $P(y=1)$:
$$P(y=1) = g(x) = \frac{1}{1+e^{-x\theta}}$$

Logistic Regression

$$P(y = 1|x) = g(x) = \frac{1}{1 + e^{-x\theta}}$$

- This function, Let $g(z) = \frac{1}{1+e^{-z}}$ is called the *sigmoid* or *logistic* function
    - Tends to 0 as z decreases
    - Tends to 1 as z increases
- This has the nice characteristic in that it's differentiable
    - Why might that be important?!



Logistic Regression

- Let $g_\theta(x) = g(x\theta) = \frac{1}{1+e^{-x\theta}}$
- Then we can compute the probabilities
  - $P(y = 1|x, \theta) = g_\theta(x)$
  - $P(y = 0|x, \theta) = 1 - g_\theta(x)$
- Which we often refer to as the *likelihoods.*
- Ultimately we want to find the parameters θ to minimize the classification error
  - Or find the parameters θ to maximize the correct class likelihood

Fit Parameters Based on Maximum Likelihood

- Given the true value y we can then compute the likelihood that we are correct as

$$P(y|x,\theta) = \left(g_\theta(x)\right)^{y}\left(1 - g_\theta(x)\right)^{(1-y)}$$

- Doing this for the entire dataset we're interested in

$$P(Y_1, Y_2, \dots, Y_N | X_1, X_2, \dots, X_N, \theta)$$

- Since $0 \leq P(Y_t | X_t, \theta) \leq 1$ we can treat them as probabilities
- Since the observations are *conditionally independent* of one another

$$P(Y|X,\theta) = P(Y_1, \dots, Y_N | X_1, \dots, X_N, \theta) = \prod_{t=1}^{N} P(Y_t | X_t, \theta)$$

- Therefore, doing this for all samples we get an overall value for our "correctness" as

$$P(Y|X,\theta) = \prod_{t=1}^{N} P(Y_t | X_t, \theta) = \prod_{t=1}^{N} \left(g_\theta(X_t)\right)^{Y_t}\left(1 - g_\theta(X_t)\right)^{(1-Y_t)}$$

Log Likelihood

- So what do we do with this likelihood $P(Y|X,\theta)$?
  - We want to maximize it!
- So we're going to want to take the derivative
- But taking the derivative of a product of a lot of things involves a very long expansion
- Let's instead first take the *log* of this
  - Doing so will result in a sum which is easier to take the derivative of.
  - So now we want to *maximize the log likelihood*

Log Likelihood

- From the properties of logarithms
  - $\log_b(mn) = \log_b(m) + \log_b(n)$
  - $\log_b(m^n) = n \cdot \log_b(m)$
- Returning to our likelihood $P(Y_t|X_t, \theta)$, for a single term we get
  - $ln\left(\left(g_\theta(X_t)\right)^{Y_t}\left(1 - g_\theta(X_t)\right)^{(1-Y_t)}\right)$
  - $= \ln(g_\theta(X_t)^{Y_t}) + \ln\left(1 - g_\theta(X_t)\right)^{1-Y_t}$
  - $= Y_t\, ln\left(g_\theta(X_t)\right) + (1 - Y_t)\ln\left(1 - g_\theta(X_t)\right)$

Log Likelihood

- Since we're taking the log of product of this for each instance we get a sum!

$$\ell(Y|X,\theta)$$
$$= \log P(Y|X,\theta) = \sum_{t=1}^{N} Y_t\, ln\big(g_\theta(X_t)\big) + (1 - Y_t)\ln\big(1 - g_\theta(X_t)\big)$$

To Maximize Likelihood

$$\ell(Y|X,\theta) = \sum_{t=1}^{N} Y_t\, ln\big(g_\theta(X_t)\big) + (1 - Y_t)\ln\big(1 - g_\theta(X_t)\big)$$

- Ideally we'd like to take the derivative of this with respect to θ, set it equal to zero, and solve for θ to find the maxima
  - The closed form approach
  - But this isn't easy ☹
- So what's our other approach
  - Do partial derivatives on the parameters and use gradient descent! (actually in this case gradient ascent, since we're trying to maximize)

To Maximum Likelihood

- We're going to take the partial derivatives with respect to θ_j
- Like before for simplicity let's start off with just one training instance (x, y)

$$\ell(y|x,\theta) = y\ln\big(g_\theta(x)\big) + (1-y)\ln\big(1-g_\theta(x)\big)$$

- Therefore we want

$$\frac{\partial}{\partial\theta_j}\ell(y|x,\theta) = \frac{\partial}{\partial\theta_j}\Big(y\ln\big(g_\theta(x)\big) + (1-y)\ln\big(1-g_\theta(x)\big)\ \Big)$$

To Maximum Likelihood

$$\frac{\partial}{\partial\theta_j}\ell(y|x,\theta) = \frac{\partial}{\partial\theta_j}\left(y\ln\left(g_\theta(x)\right) + (1-y)\ln\left(1-g_\theta(x)\right)\ \right)$$

- We'll need the partial of the sigmoid, $g_\theta(x)$ with respect to θ_j:
  - $\frac{\partial}{\partial\theta_j} g_\theta(x) = \frac{\partial}{\partial\theta_j}\left(\frac{1}{1+e^{-x\theta}}\right) = \frac{\partial}{\partial\theta_j}\left(1+e^{-x\theta}\right)^{-1}$
  - $= -1\left(0 + x_j e^{-x\theta}\right)\left(1+e^{-x\theta}\right)^{-2} = -\frac{x_j e^{-x\theta}}{\left(1+e^{-x\theta}\right)^2}$
  - $= -\frac{1}{1+e^{-x\theta}}\frac{e^{-x\theta}}{1+e^{-x\theta}}x_j$
  - $= g_\theta(x)\left(1-g_\theta(x)\right)x_j$

To Maximum Likelihood

$$\frac{\partial}{\partial \theta_j} \ell(y|x,\theta) = \frac{\partial}{\partial \theta_j} \left(y \ln\big(g_\theta(x)\big) + (1-y) \ln\big(1 - g_\theta(x)\big) \right)$$

- From the previous slide we have

$$\frac{\partial}{\partial \theta_j} g_\theta(x) = g_\theta(x)\big(1 - g_\theta(x)\big)x_j$$

- Derivation in class…
  - To help $\frac{\partial}{\partial x} \ln x = \frac{1}{x} \cdot \frac{\partial}{\partial x} x$
- Results in:

$$\frac{\partial}{\partial \theta_j} \ell(y|x,\theta) = \big(y - g_\theta(x)\big)x_j$$

Gradient Ascent Rule

$$\frac{\partial}{\partial \theta_j} \ell(y|x,\theta) = \big(y - g_\theta(x)\big)x_j$$

- We want this to go towards zero (local maxima)
- So let's update θ_j as

$$\theta_j \coloneqq \theta_j + \eta \frac{\partial}{\partial \theta_j} \ell(y|x,\theta)$$

$$\theta_j = \theta_j + \eta\big(y - g_\theta(x)\big)x_j$$

- This is the same form as the sum-of-squares error for linear regression!!!!

Logistic Regression Example

- Let's classifying whether a person will buy a product or not

	Y	X-Variables							
Obs. No.	Buy	Income	Is Female	Is Married	Has College	Is Pro-fessional	(Omit-ted Vari-ables)	Prev Child Mag	Prev Parent Mag
1	0	24000	1	0	1	1	...	0	0
2	1	75000	1	1	1	1	...	1	0
3	0	46000	1	1	0	0	...	0	0
4	1	70000	0	1	0	1	...	1	0
5	0	43000	1	0	0	0	...	0	1
6	0	24000	1	1	0	0	...	0	0
7	0	26000	1	1	1	0	...	0	0
8	0	38000	1	1	0	0	...	0	0
9	0	39000	1	0	1	1	...	0	0
10	0	49000	0	1	0	0	...	0	0
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
654	0	10000	1	0	0	0	...	0	0
655	1	75000	0	1	0	1	...	0	0
656	0	72000	0	0	1	0	...	0	0
657	0	33000	0	0	0	0	...	0	0
658	0	58000	0	1	1	1	...	0	0
659	1	49000	1	1	0	0	...	0	0
660	0	27000	1	1	0	0	...	0	0
661	0	4000	1	0	0	0	...	0	0
662	0	40000	1	0	1	1	...	0	0
663	0	75000	1	1	1	0	...	0	0
664	0	27000	1	0	0	0	...	0	0
665	0	22000	0	0	0	1	...	0	0
666	0	8000	1	1	0	0	...	0	0
667	1	75000	1	1	1	0	...	0	0
668	0	21000	0	1	0	0	...	0	0
669	0	27000	1	0	0	0	...	0	0
670	0	3000	1	0	0	0	...	0	0
671	1	75000	1	1	0	1	...	0	0
672	1	51000	1	1	0	1	...	0	0
673	0	11000	0	1	0	0	...	0	0

KidCreative.csv

Logistic Regression Example

- Make some design decisions:
  - Randomize data
  - Use 2/3 training, 1/3 testing
  - Standardize features
  - Add bias feature
  - Initialize parameters to random values in the range [-1, 1]
  - Since our equation is based on log likelihood, let's terminate when change in sum of log likelihoods doesn't change more than eps
    - Recall the log likelihood of an example being correct is
    $$y\ln\left(\frac{1}{1+\mathrm{e}^{-x\theta}}\right)+(1-y)\ln\left(1-\frac{1}{1+\mathrm{e}^{-x\theta}}\right)$$
    - But be careful… $\log(0)=-Inf$. So you might need to deal with this somehow
  - Let's do something different (smarter?) with η
    - Start if off relatively large, say $\eta=0.5$
    - If we notice an decrease in the log likelihood (meaning we over-jumped the maxima), then decrease it by ½
  - Let's do batch regression

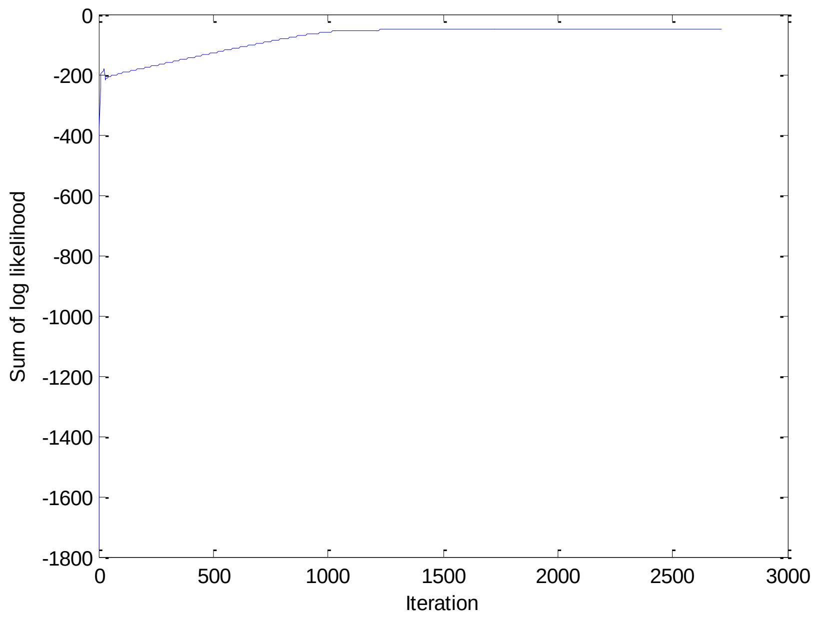
Example

θ =

-12.7663
5.1987
0.9057
0.6136
0.0880
-0.3606
-0.4920
-29.3976
0.2140
-0.0329
0.7355
0.3957
0.0054
0.9061
1.1647
0.1887
0.1881

Choosing Class 1 if $P(y = 1|x, \eta) \geq 0.5$ we get:

Precision: 0.7708
Recall: 0.7551
F-Measure: 0.7629



Final Observations

- Let’s think about this algorithm
  - Supervised or non?
  - Classification or regression?
  - Model-based or instance-based?
    - When it comes time to test/use, are we using the original data?
  - Linear vs Non-Linear?
  - Can this work on categorical data?
  - Can this work on continuous valued data?
  - Training Complexity?
  - Testing Complexity?
  - How to deal with overfitting?
  - Directly handles multi-class?