CPSC 340: Machine Learning and Data Mining

L1-Regularization

… and …

Maximum likelihood

BONUS SLIDES

Bonus Slide: L2- vs. L1-Regularization for Sparsity

Influence of L2 and L1 regularization on sparsity

I am bit confused about how the L1 and L2 norm affect sparsity of parameters. From what I understand about regularization, they both penalize non-zero parameters and so my gut instinct tells me that they would both encourage sparsity as we increase lambda. However, it seems that L1 regularization encourages sparsity while L2 regularization does not. I am super confused by this and would appreciate any help! Thanks!

hw5

edit · good question | 1

the instructors' answer, *where instructors collectively construct a single answer*

They both encourage variables to move closer to zero. But L2-regularization does not encourage variables to be exactly zero while L1-regularization encourages variables to be exactly 0.

To see why, think about the penalty we apply to very small values w_j. For example, if $w_j = 0.0001$ then because we square it the L2-penalty on this element will be $\lambda 0.00000001$. The closer w_j gets to zero, the smaller the L2-penalty gets. In contrast, the L1-penalty always penalizes by $\lambda 0.0001$ and the effect of the penalty stays always proportional to $|w_j|$ as w_j approaches zero.

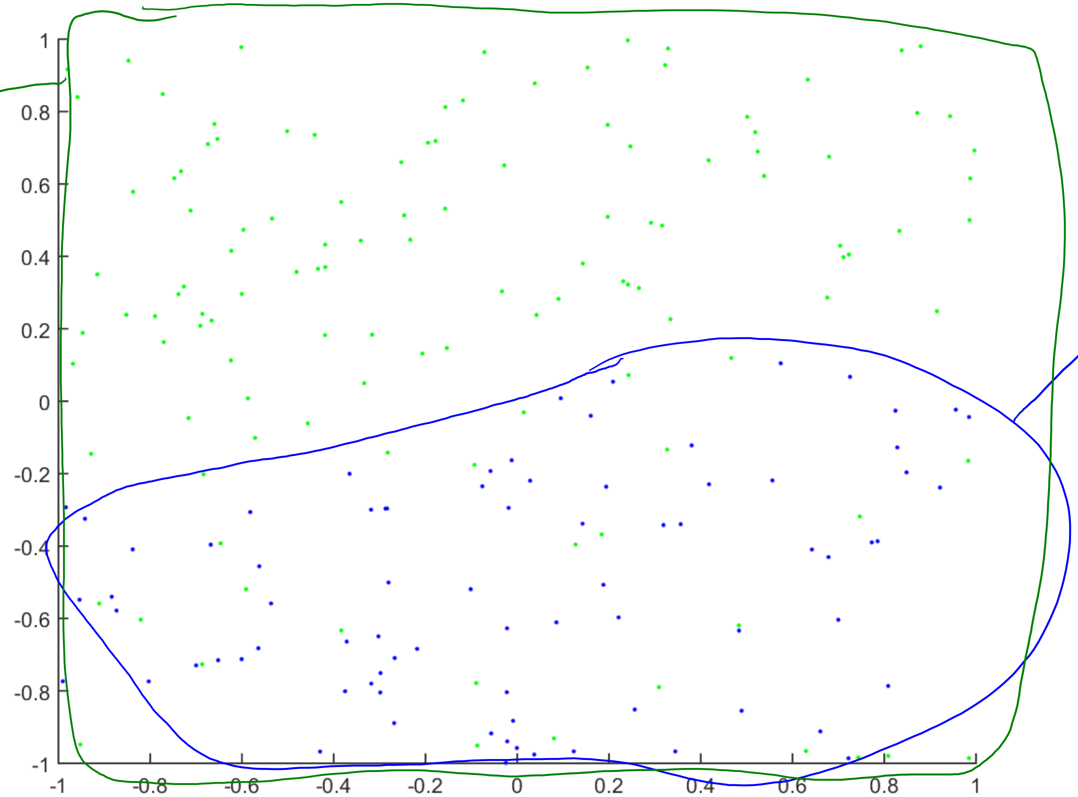
Bonus Slide: Other Parsimonious Parameterizations

- Sigmoid isn't the only parsimonious $p(y_i \mid x_i, w)$:
  - Noisy-Or (simplier to specific probabilities by hand).
  - Probit (uses CDF of normal distribution, very similar to logistic).
  - Extreme-value loss (good with class imbalance).
  - Cauchit, Gosset, and many others exist…

Bonus Slide: Unbalanced Data and Extreme-Value Loss

- Consider binary case where:
  - One class overwhelms the other class ('unbalanced' data).
  - Really important to find the minority class (e.g., minority class is tumor).

"majority" class is everywhere.



important "minority" class

Bonus Slide: Unbalanced Data and Extreme-Value Loss

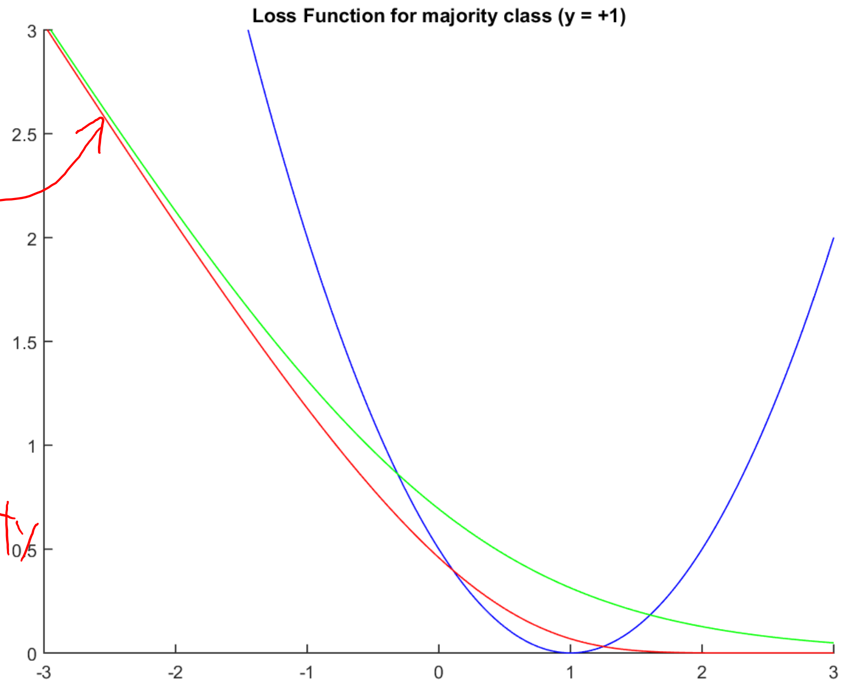
- Extreme-value distribution:

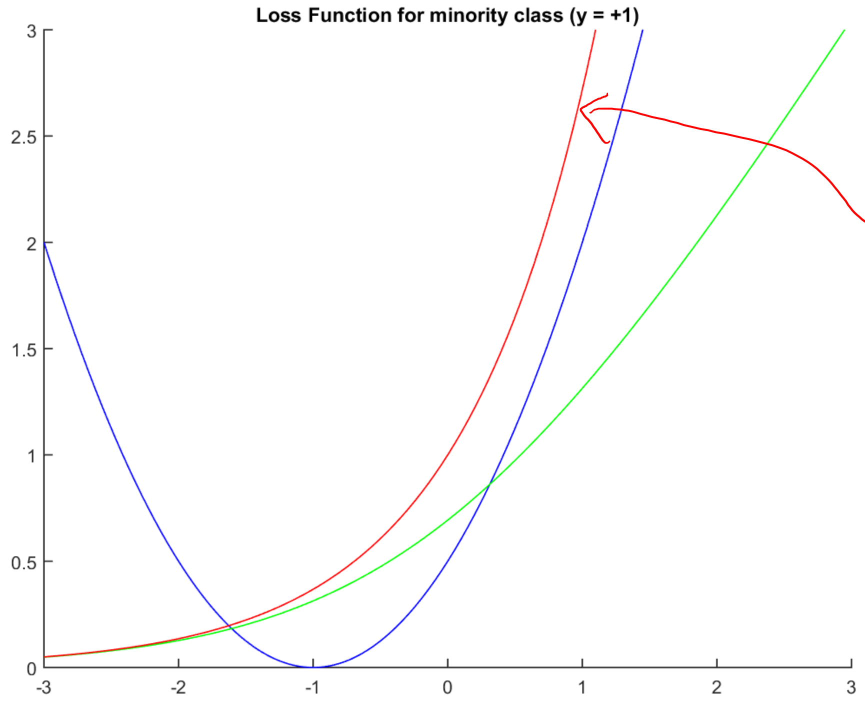
$$p(y_i=+1\,|\,\hat{y}_i) = 1-\exp(-\exp(\hat{y}_i)) \quad [+1 \text{ is majority class}]$$

To make it a probability, $p(y_i=-1\,|\,\hat{y}_i) = \exp(-\exp(\hat{y}_i))$

asymmetric

Similar to logistic for majority class





big penalty for getting minority class wrong.

Bonus Slide: Unbalanced Data and Extreme-Value Loss

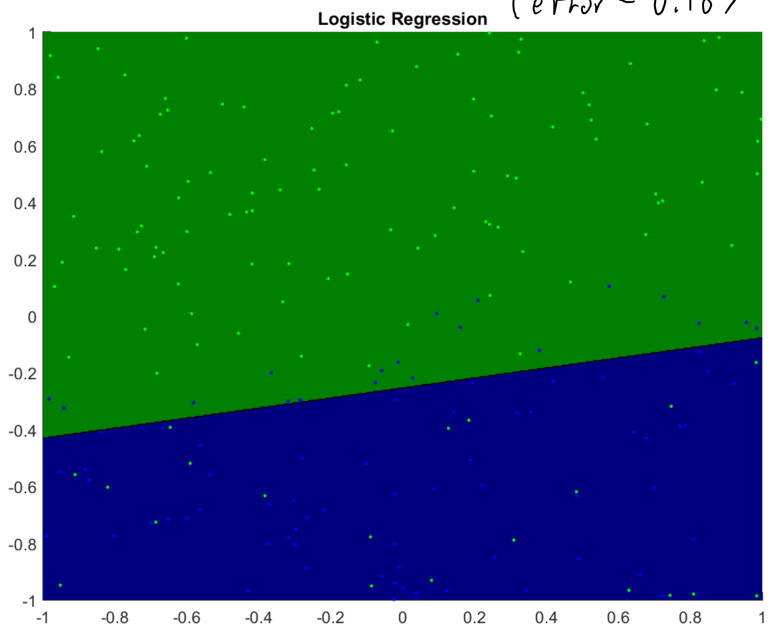
- Extreme-value distribution:

$$p(y_i = +1 \mid \hat{y}_i) = 1 - \exp(-\exp(\hat{y}_i)) \quad [+1 \text{ is majority class}]$$

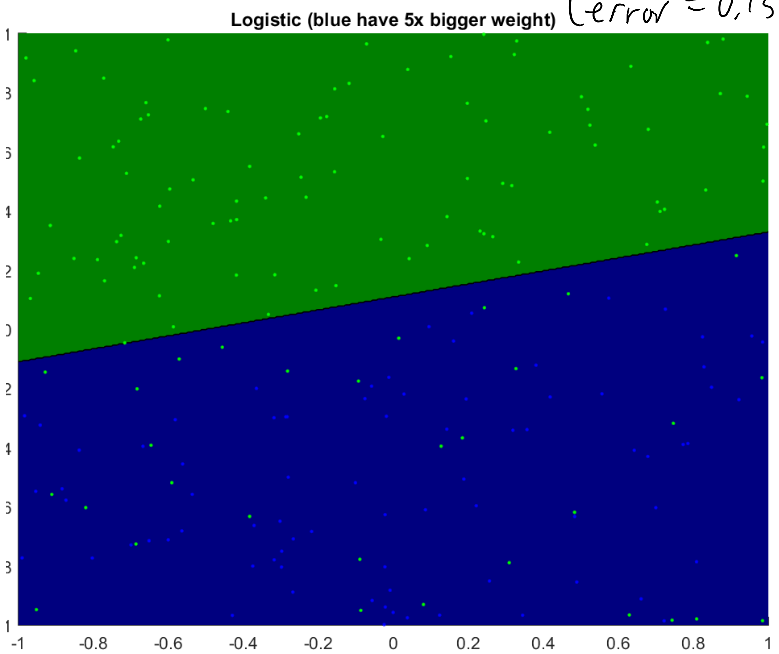
asymmetric

To make it a probability, $p(y_i = -1 \mid \hat{y}_i) = \exp(-\exp(\hat{y}_i))$

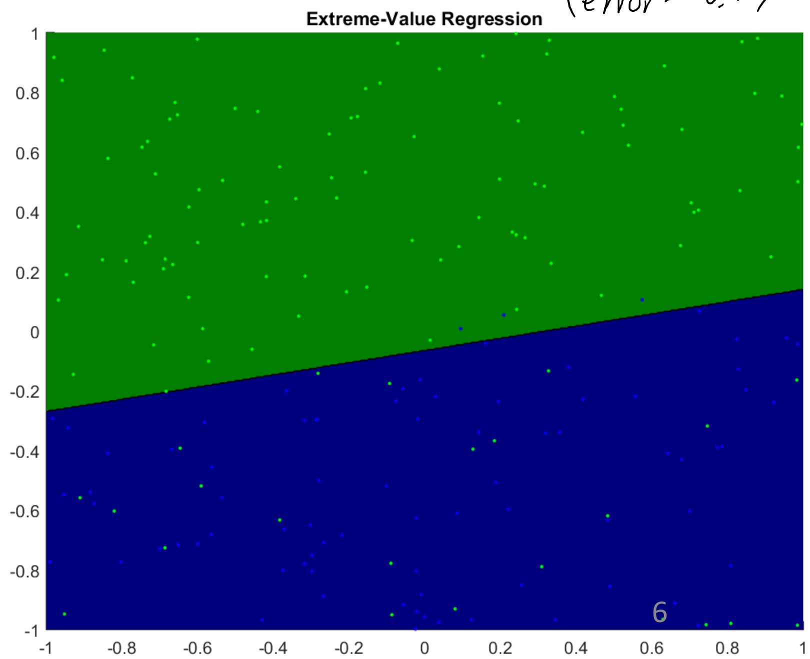
(error = 0.18)



(error = 0.15)



(error = 0.13)



Bonus Slide: “Heavy” Tails vs. “Light” Tails