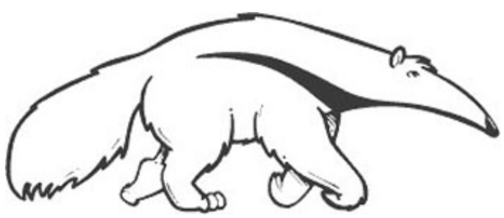
Machine Learning and Data Mining

Loss Functions

Prof. Alexander Ihler

Fall 2012



BREN:ICS INFORMATION AND COMPUTER SCIENCES

UNIVERSITY of CALIFORNIA IRVINE

Loss functions

- Measure error in our predictions
  - A function of the parameters and training data
- $J(\theta) = \ldots$

- Ideally, these should
  - Measure what we care about
  - Be easy to optimize over

- Often these two goals are in conflict…

Cost functions for regression

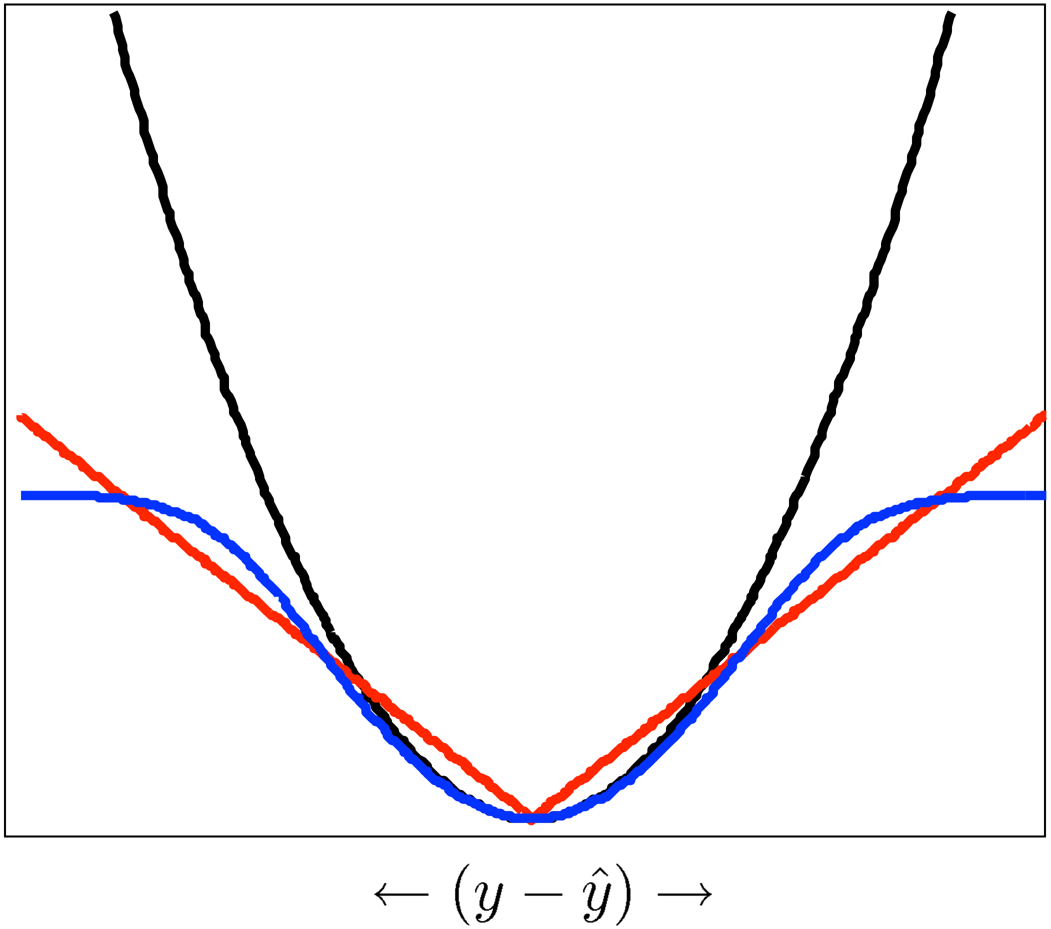
$\ell_2 \;:\; (y - \hat{y})^2$ **(MSE)**

$\ell_1 \;:\; |y - \hat{y}|$ **(MAE)**

Something else entirely…

$c - \log(\exp(-(y - \hat{y})^2) + c)$

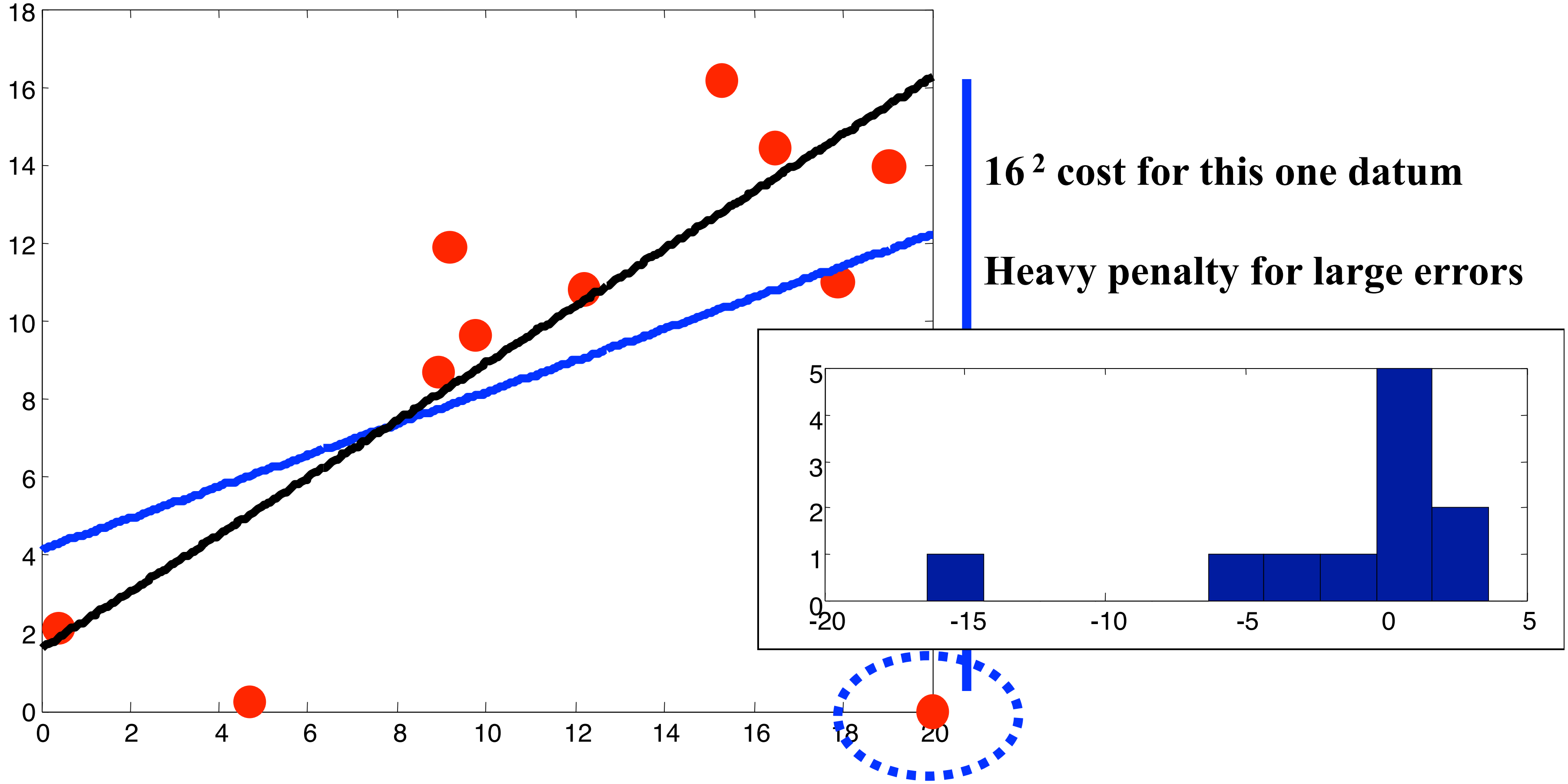
(???)



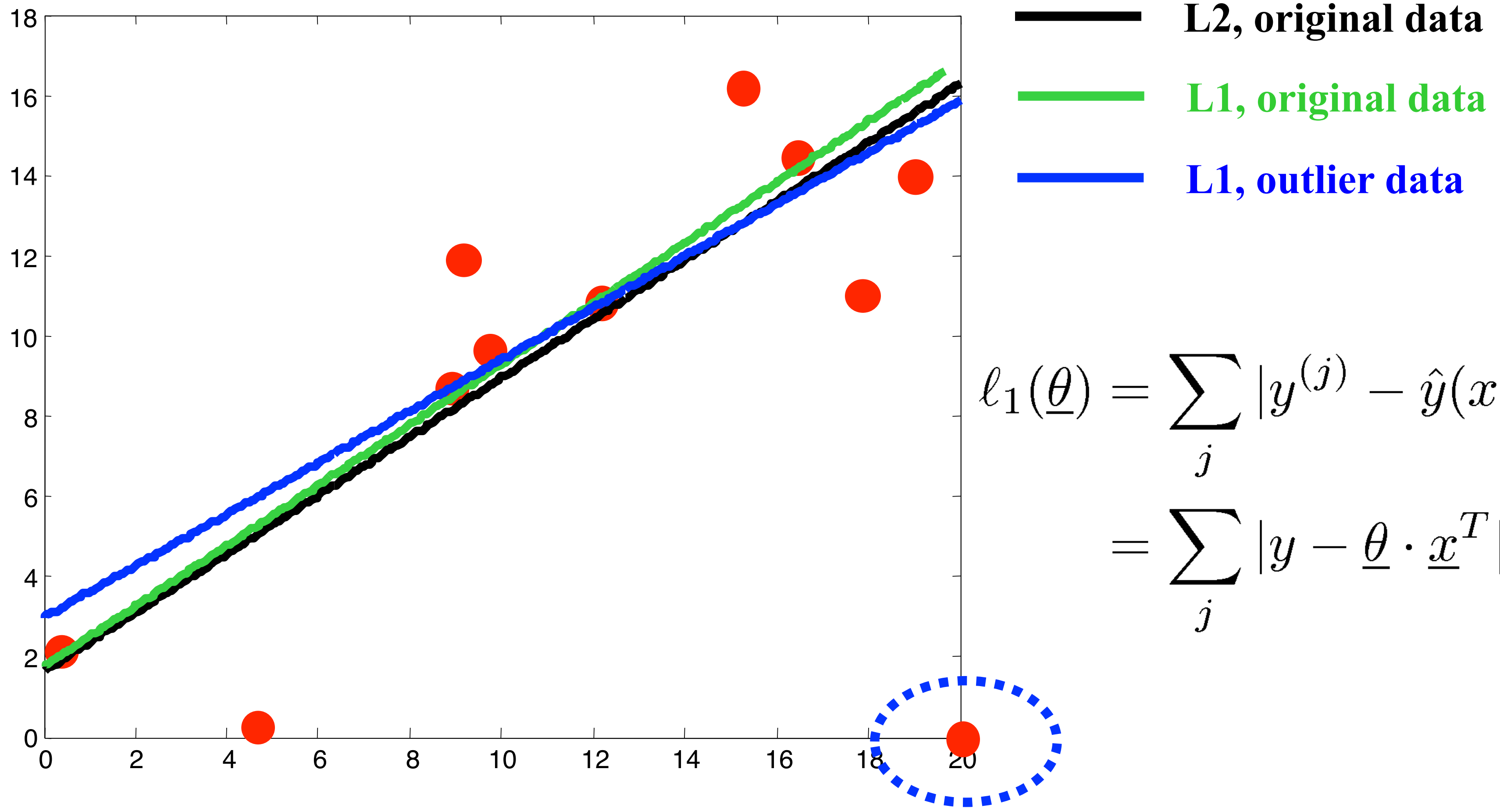
"Arbitrary" functions can't be solved in closed form…
- use gradient descent

Effects of cost function choice

- Sensitivity to outliers



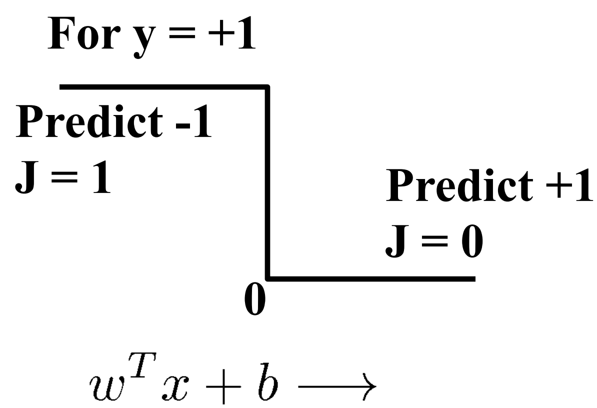
L1 error



$$\ell_1(\underline{\theta}) = \sum_j |y^{(j)} - \hat{y}(x^{(j)})|$$
$$= \sum_j |y - \underline{\theta} \cdot \underline{x}^T|$$

Classification cost functions

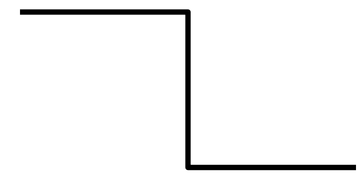
- Consider a linear classifier
- J(.) = # of misclassified data?
  - Not smooth = hard to train



- This is called the 0/1 loss
  - Cost 0 when we're right; cost 1 when we're wrong
- Often, it's what we care about
  - Measures the number of mistakes we will make
- It's hard to optimize
  - No incentive to be "less wrong" or "more right"

Surrogate loss functions

- Replace 0/1 loss $J(\theta, x^{(i)}) = \delta\big(T(\theta x^{(i)}) \neq y^{(i)}\big)$ with something easier:

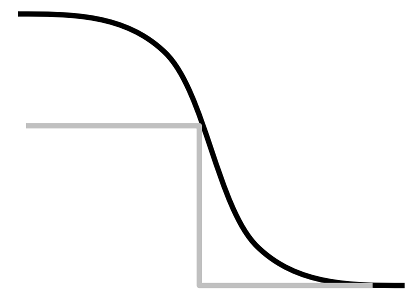


0/1 Loss

y={-1 , +1}

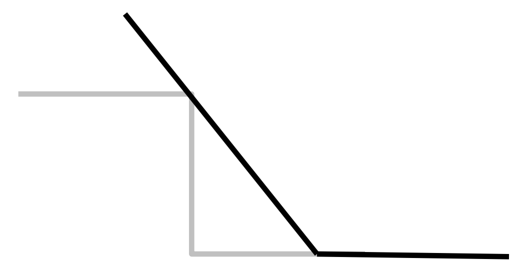
- Logistic MSE

$$J(\theta, x^{(i)}) = \big(\sigma(\theta x^{(i)}) - y^{(i)}\big)^2$$



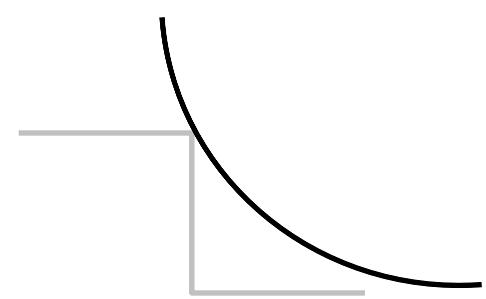
- Hinge loss

$$J(\theta, x^{(i)}) = \max\left[0\,,\,1 - y^{(i)}\,\theta x^{(i)}\right]$$



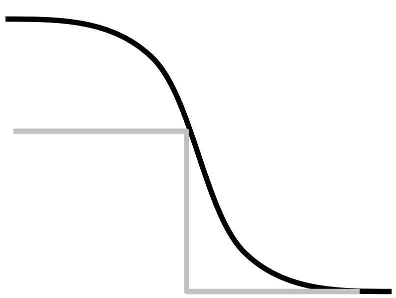
- Exponential loss

$$J(\theta, x^{(i)}) = \exp\left[-y^{(i)}\,\theta x^{(i)}\right]$$

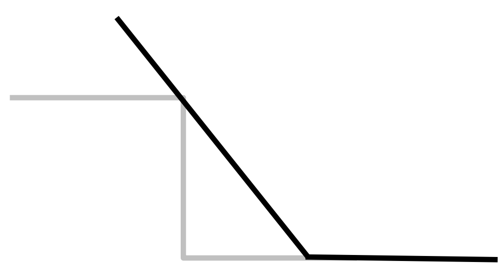


Surrogate loss functions

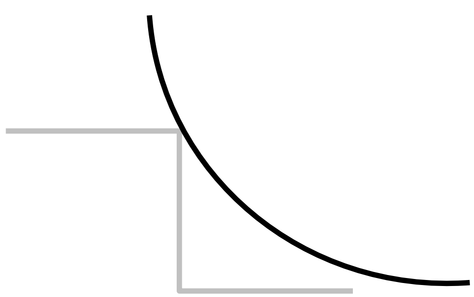
- Properties of a good loss function
  - Close to desired "real" loss?
    - Upper bound: low surrogate loss => low real loss
  - Smooth
  - Derivative = 0 only if real cost = 0
  - Convex?
    - Easy to optimize; no local optima



Logistic MSE



Hinge



Exponential