

Introduction to Deep Learning

5. Maximum Likelihood and Logistic Regression

STAT 157, Spring 2019, UC Berkeley

Mu Li and Alex Smola

courses.d2l.ai/berkeley-stat-157



MAGIC Etch A Sketch[®] SCREEN

Logistics Update



Horizontal
Dial



Vertical
Dial

OHIO ART  The World of Toys[®]

MAGIC SCREEN IS GLASS SET IN STURDY PLASTIC FRAME
USE WITH CARE

courses.

aws 

Homework

- Please follow the submission instructions on HW3
 - Don't submit `homework1_v1` `homework1_a` `homework1_updated` and hope that we figure it out ...
 - Please check the PDF that it is readable
- Homework 1 solutions have an example for how to format
- Converting notebooks to PDF
`ipython nbconvert --to pdf notebook.ipynb`

Project

- If you haven't found a team yet, do it TODAY
- Email to berkeley-stat-157@googlegroups.com with
 - Names
 - Project title
 - A short abstract (optional)
- Extended deadline - February 6, 11:59 PM PST

Outline

- **Maximum Likelihood**
 - Gauss and means
 - More loss functions (l_1 loss, trimmed mean)
 - Regression revisited
- **Classification**
 - Computing discrete probabilities
 - Likelihood and loss functions
- **Information Theory**



MAGIC Etch A Sketch[®] SCREEN

Maximum
Likelihood
 \neq MAP



Horizontal
Dial

OHIO ART  The World of Toys[®]

Vertical
Dial



MAGIC SCREEN IS GLASS SET IN STURDY PLASTIC FRAME
DO NOT USE WITH CARE



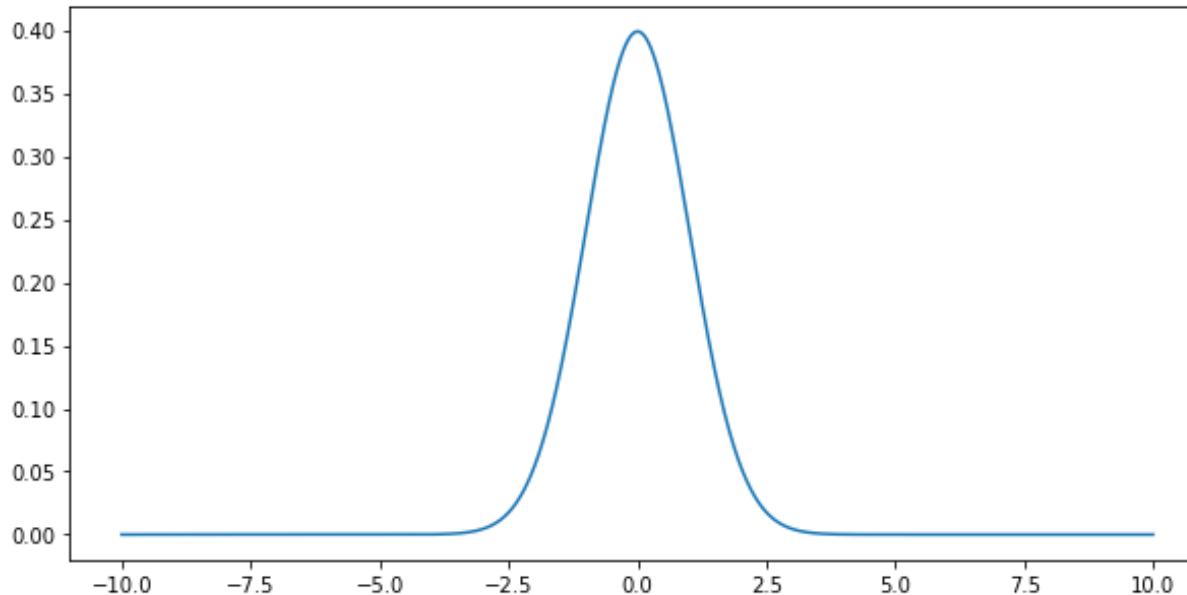
courses.

aws 

Flashback - Normal Distribution

Density

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Estimating the parameters in a Gaussian

- **Mean**

$$\mu = \mathbf{E}[x] \text{ hence } \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

- **Variance**

$$\sigma^2 = \mathbf{E}[(x - \mu)^2] \text{ hence } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

Why?

Likelihood

- Observe some data $X = \{x_1, \dots, x_n\}$
- Assume that the data is drawn from a Gaussian

$$p(X; \mu, \sigma^2) = \prod_{i=1}^n p(x_i; \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

- **Fitting parameters is maximizing** $p(X; \mu, \sigma^2)$ **wrt.** μ, σ^2
(maximize likelihood that data was generated by model)
- **Practical simplification**

$$\underset{\mu, \sigma^2}{\text{maximize}} p(X; \mu, \sigma^2) \iff \underset{\mu, \sigma^2}{\text{minimize}} -\log p(X; \mu, \sigma^2)$$

Maximum Likelihood

- Estimate parameters by finding ones that explain the data

$$\underset{\mu, \sigma^2}{\text{minimize}} -\log p(X; \mu, \sigma^2)$$

- **Decompose likelihood**

$$-\log p(X; \mu, \sigma^2) = \sum_{i=1}^n \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(x_i - \mu)^2 = \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

Minimized for $\mu = \frac{1}{n} \sum_{i=1}^n x_i$

Maximum Likelihood

- Estimating the variance

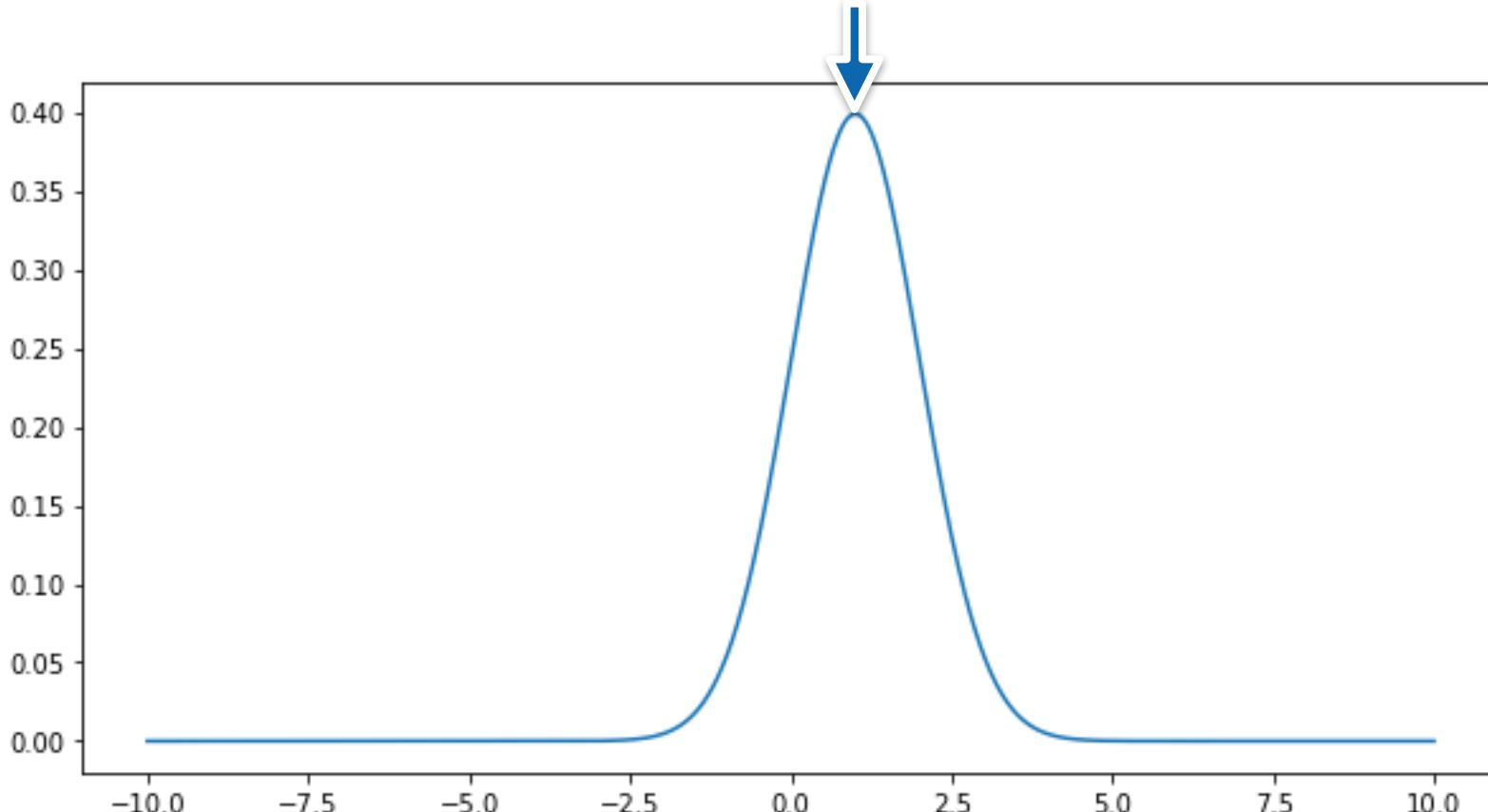
$$\frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

- Take derivatives with respect to it

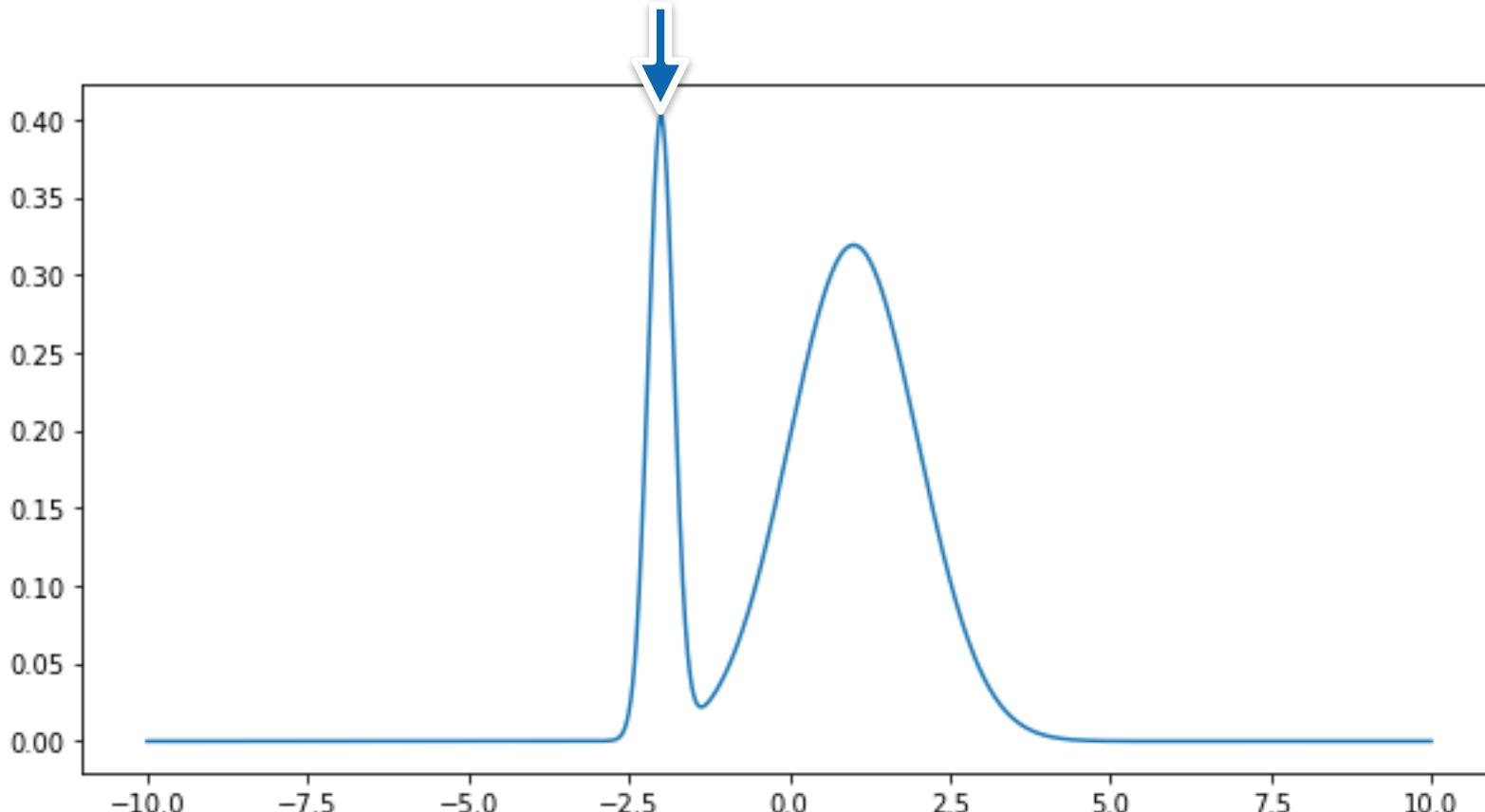
$$\partial_{\sigma^2} [\cdot] = \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

$$\implies \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

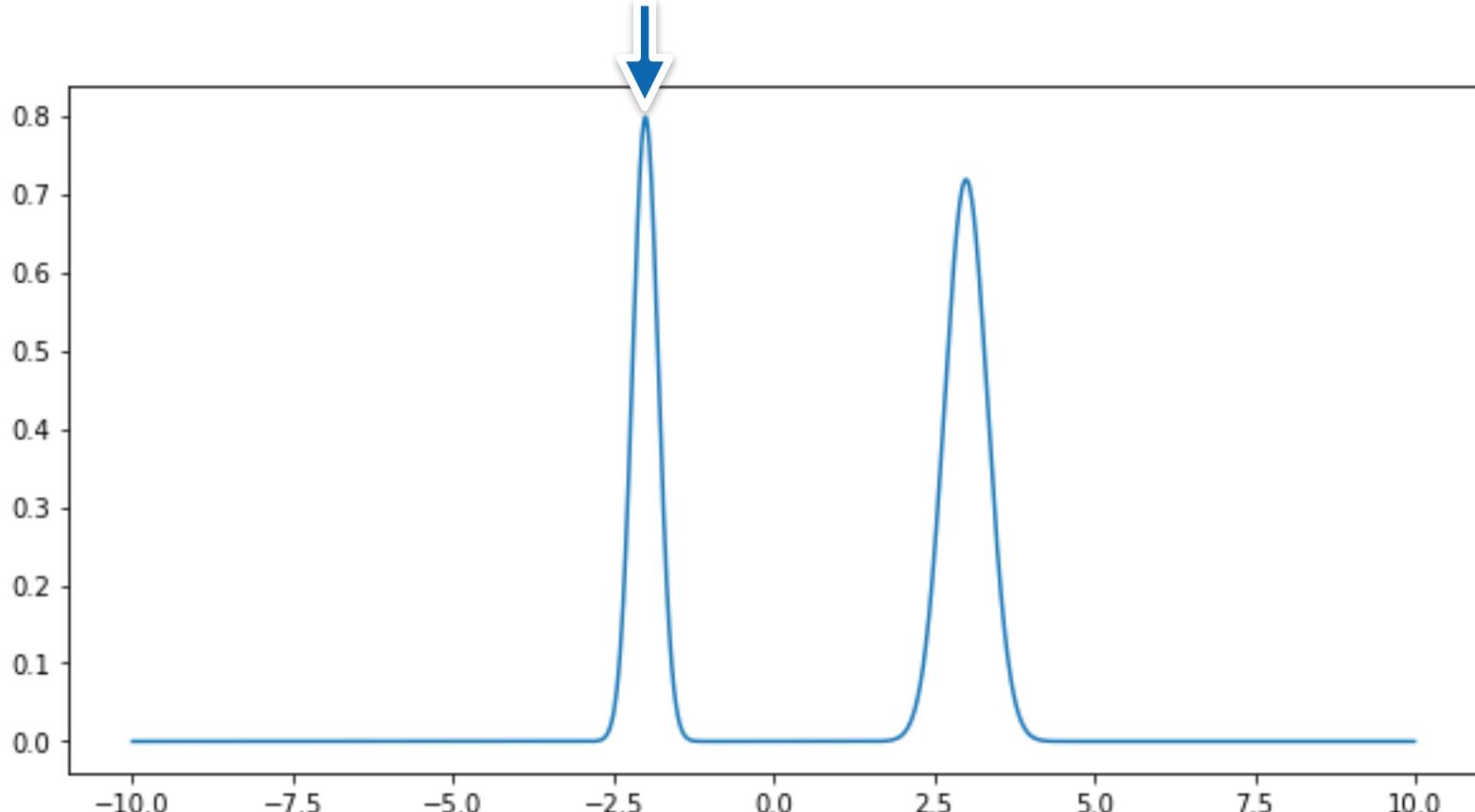
Maximum likelihood estimation



Maximum likelihood estimation



Maximum likelihood estimation



Maximum likelihood estimation

- Data - ‘student didn’t do homework’
 - Possible parameters
 - ‘dog ate homework’
 - ‘abducted by aliens’
 - ‘too lazy’
 - ‘sick grandmother’
 - All parameters explain the data.
- 

Maximum a posteriori estimation

- Posterior Probability

$$p(w | X) \propto p(X | w)p(w)$$

penalty

$$\text{hence } -\log p(w | X) = -\log p(X | w) - \log p(w) + c$$

- Maximum a Posteriori Estimation

$$\underset{w}{\text{minimize}} -\log p(X; \mu, \sigma^2) - \log p(w)$$

- No homework example

$$p(\text{'no homework'} | \text{explanation}) = 1 \text{ (all explanations work)}$$

lazy student	grandma sick	dog ate it	alien abduction
0.8	0.19	0.0099	0.0001

What does this have to do with regression?

Regression

- Recall optimization problem

$$\underset{w}{\text{minimize}} \sum_{i=1}^n (y_i - f(x_i, w))^2 + \text{penalty}(w)$$

$-\log p(w)$

Does the model work?

Additive Gaussian Noise

- Data generation model

$$y_i = f(x_i, w) + \epsilon_i \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Gaussian Prior $p(w)$ hence $-\log p(w) = \frac{1}{2\bar{\sigma}^2} \|w\|^2 + \text{const.}$

Regression

- Maximum a posteriori

$$\underset{w}{\text{minimize}} -\log p(w | X, Y)$$

$$\iff \underset{w}{\text{minimize}} \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f(x_i, w))^2 + \frac{1}{2\bar{\sigma}^2} \|w\|^2 + \text{const.}$$

$$\iff \underset{w}{\text{minimize}} \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i, w))^2 + \frac{\lambda}{2} \|w\|^2$$

Implement this



MAGIC Etch A Sketch[®] SCREEN

LOSS FUNCTIONS



Horizontal
Dial



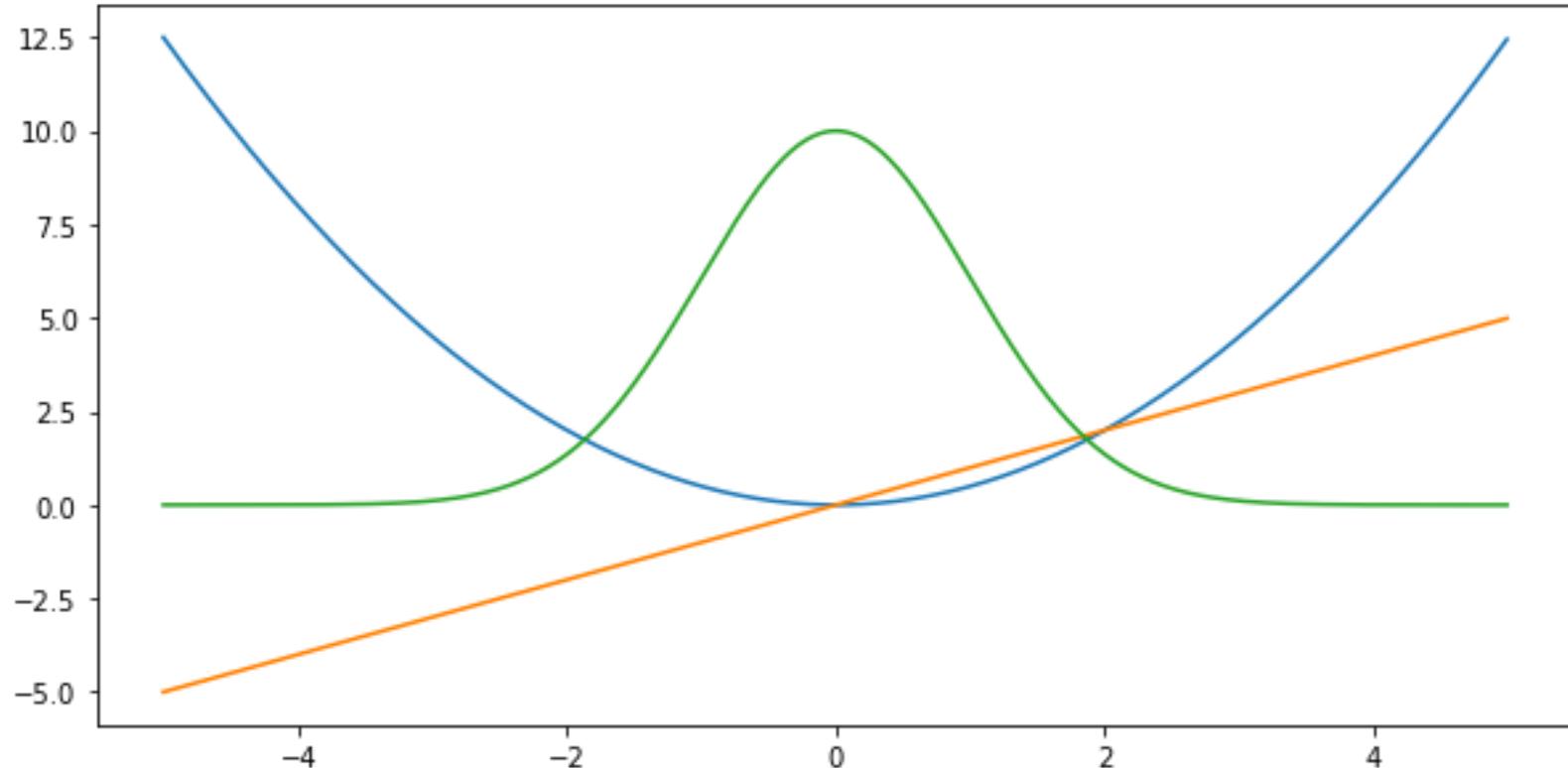
Vertical
Dial

OHIO ART  The World of Toys[®]

MAGIC SCREEN IS GLASS SET IN STURDY PLASTIC FRAME
DO NOT USE WITH CARE

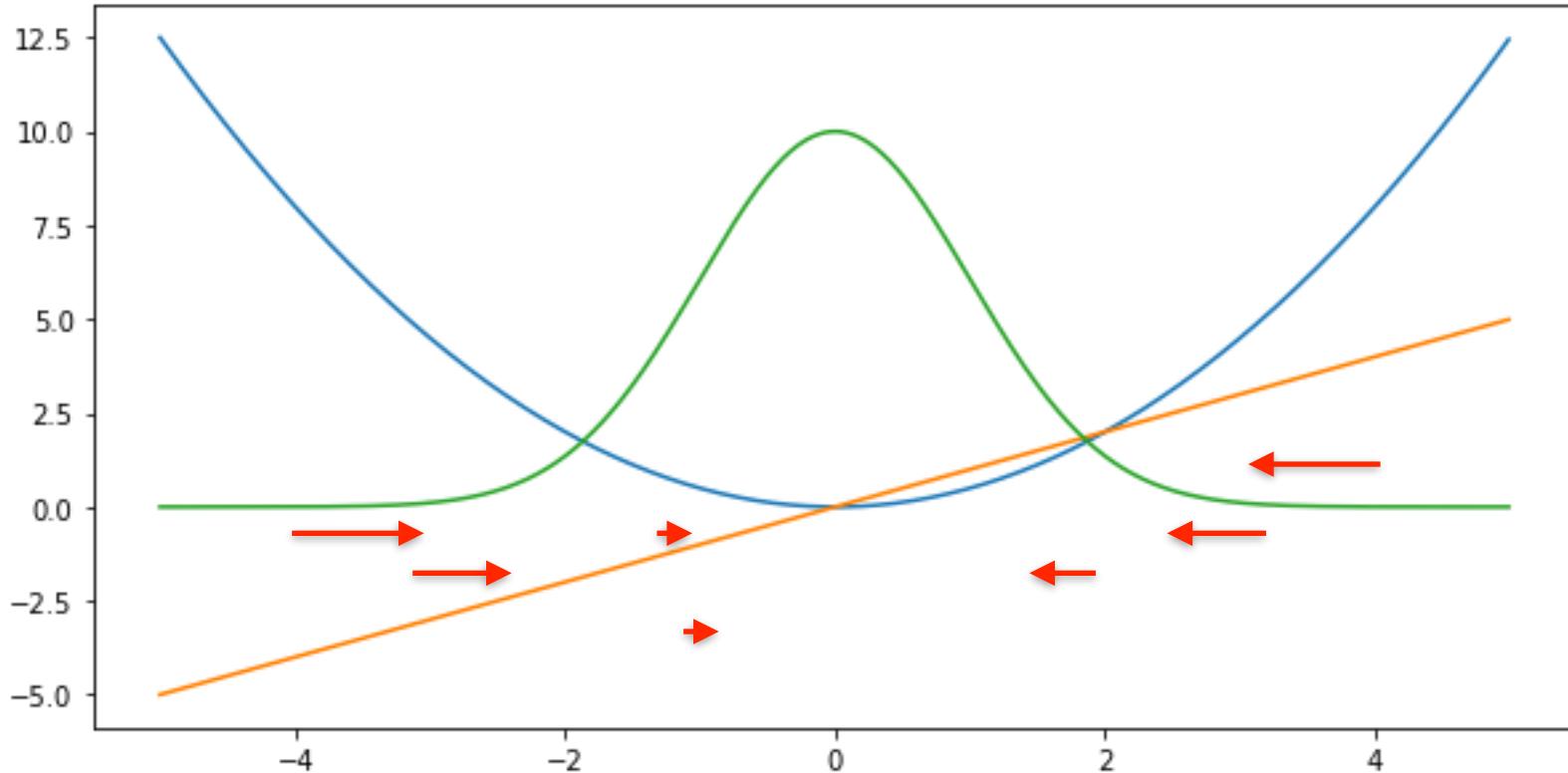
L2 Loss

$$l(y, y') = \frac{1}{2}(y - y')^2$$



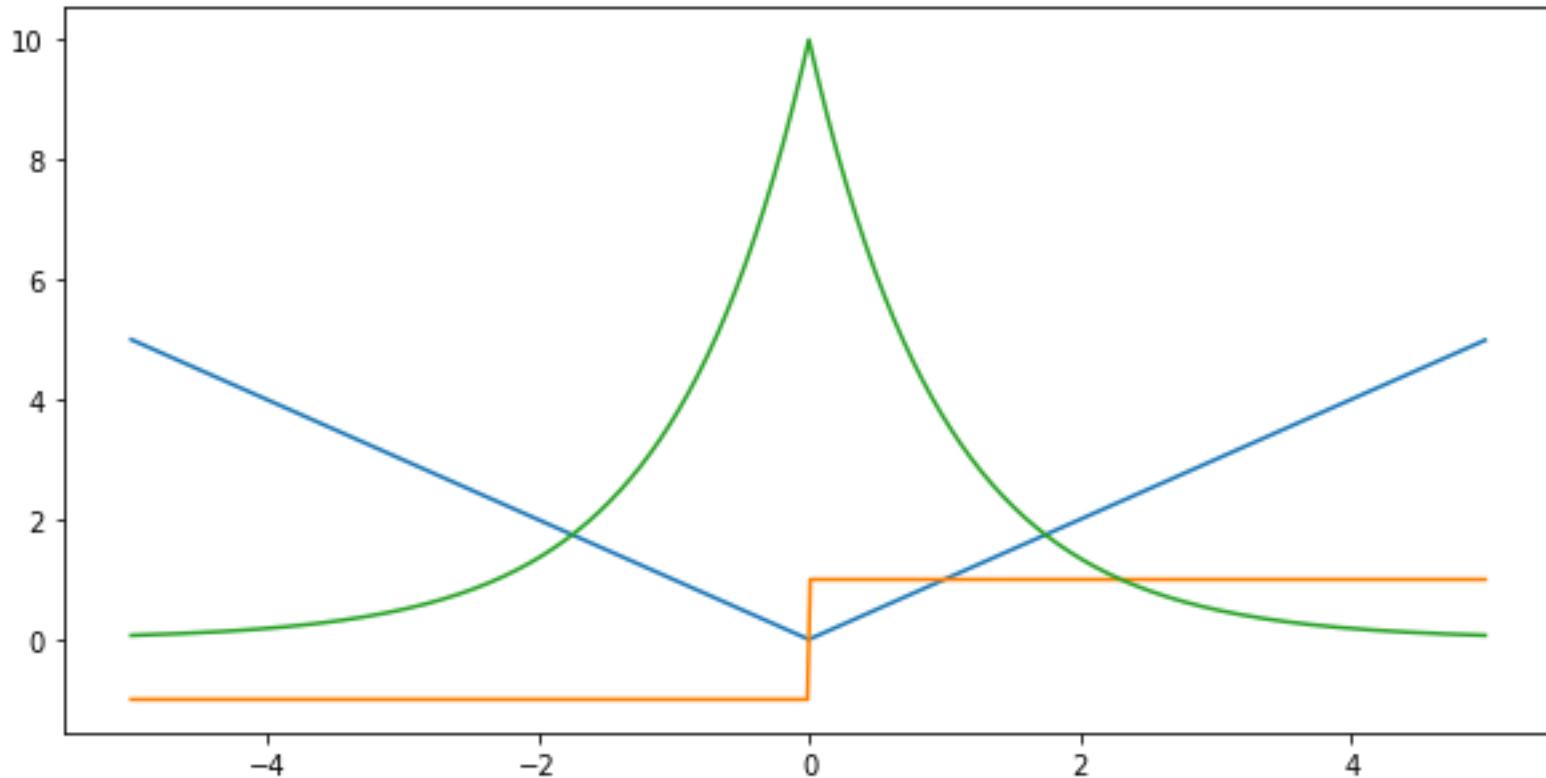
L2 Loss - mean

$$l(y, y') = \frac{1}{2}(y - y')^2$$



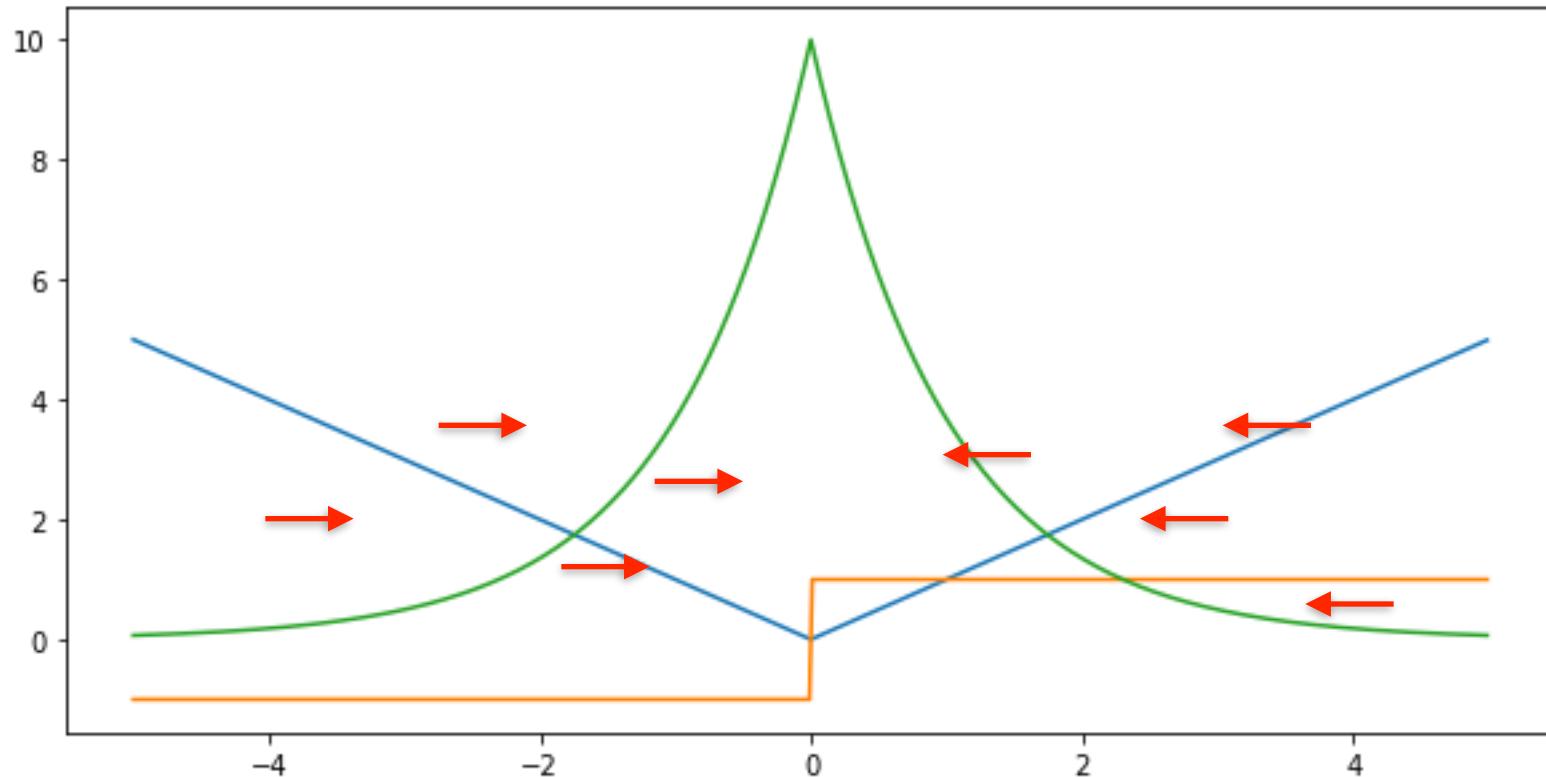
L1 Loss

$$l(y, y') = |y - y'|$$



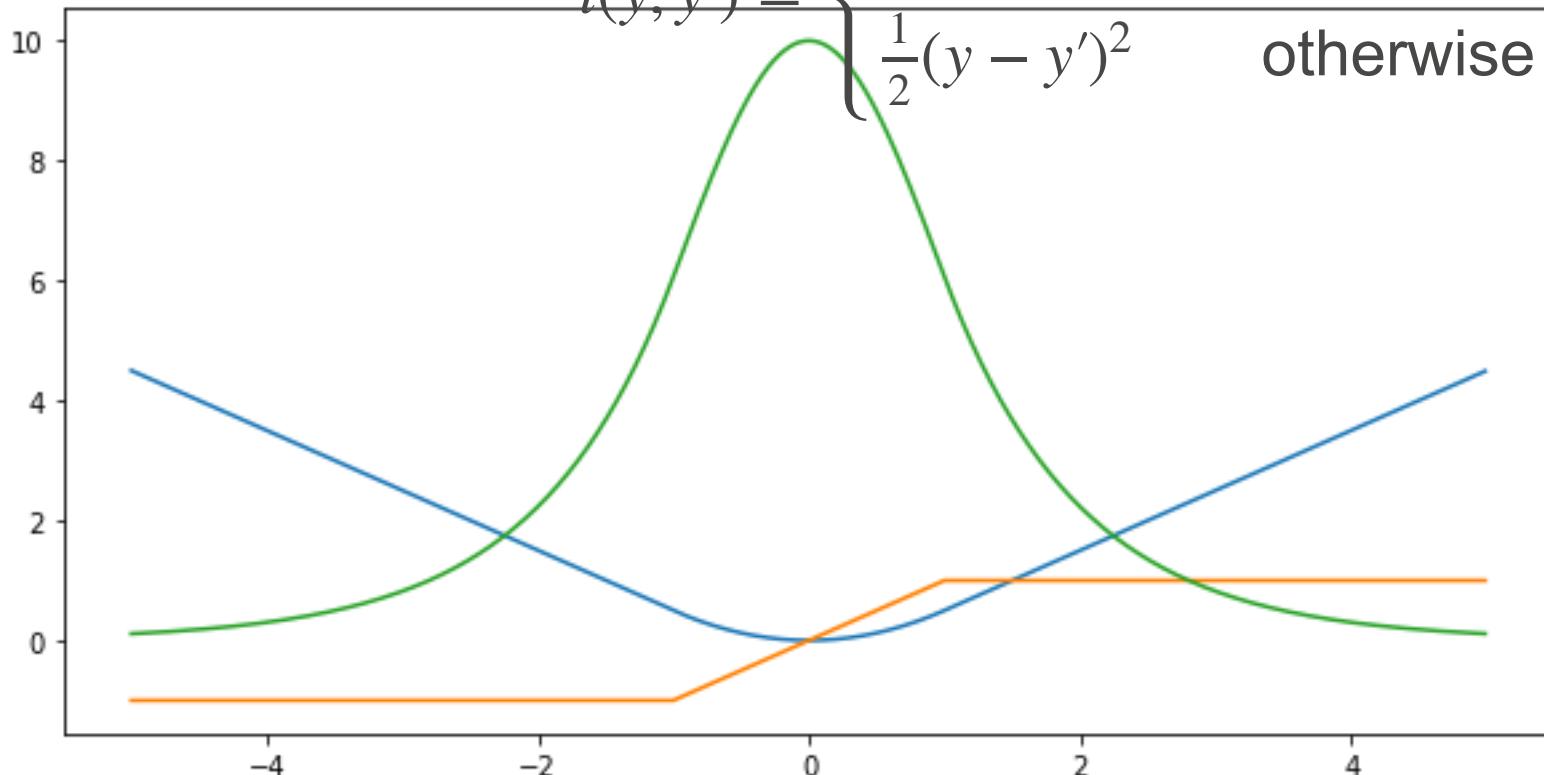
L1 Loss - median

$$l(y, y') = |y - y'|$$

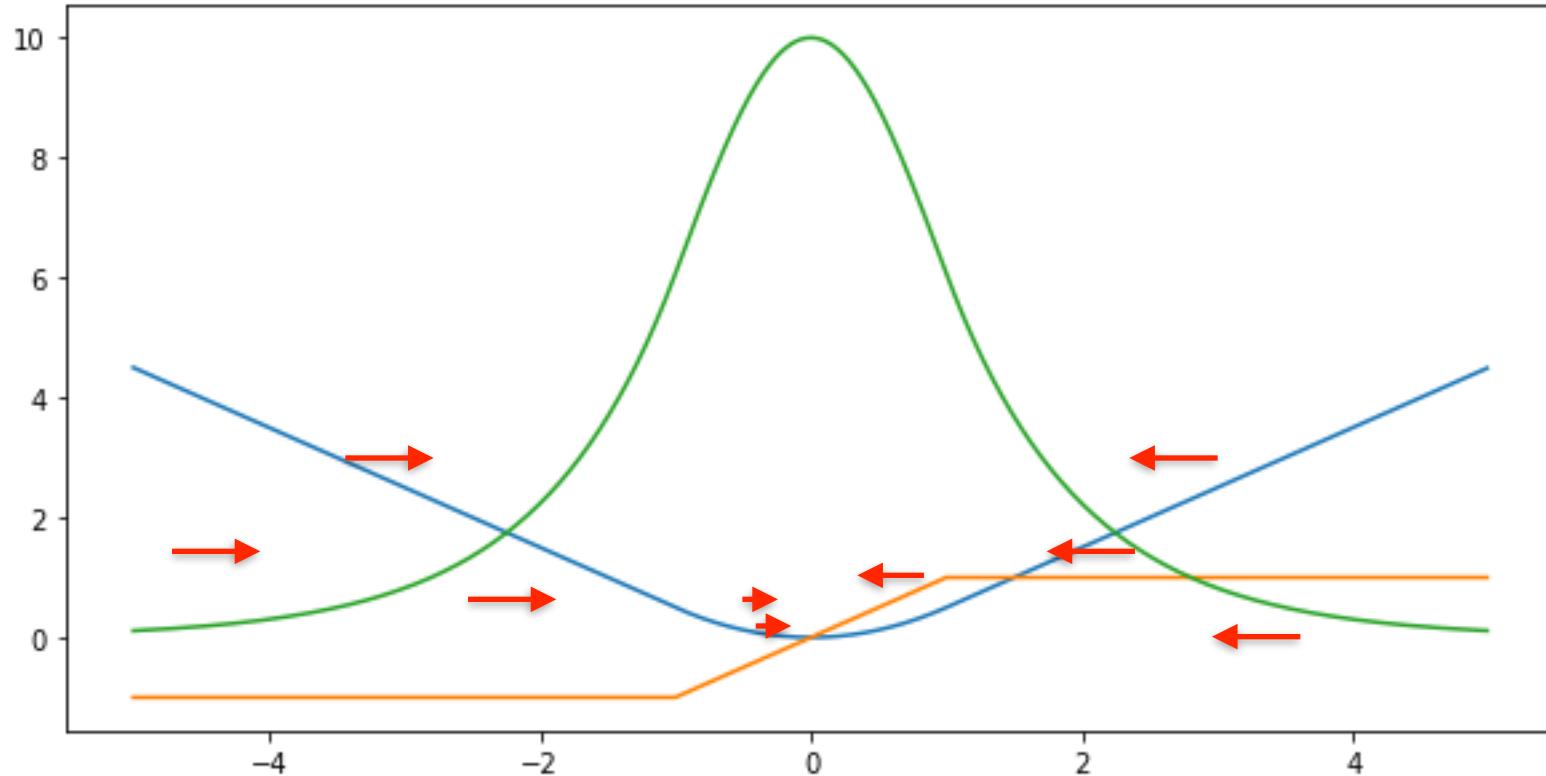


Huber's Robust Loss

$$l(y, y') = \begin{cases} |y - y'| - \frac{1}{2} & \text{if } |y - y'| > 1 \\ \frac{1}{2}(y - y')^2 & \text{otherwise} \end{cases}$$



Huber's Robust Loss - trimmed mean





MAGIC Etch A Sketch[®] SCREEN

Logistic Regression



Horizontal
Dial



Vertical
Dial

OHIO ART  The World of Toys[®]

MAGIC SCREEN IS GLASS SET IN STURDY PLASTIC FRAME
DO NOT USE WITH CARE

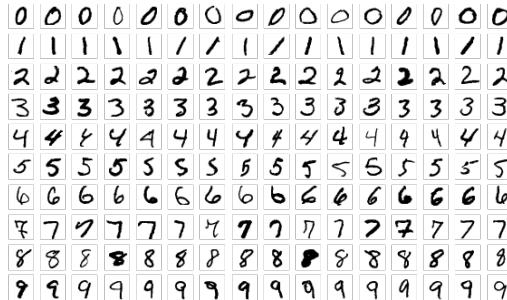
courses.

aws 

Regression vs. Classification

- Regression estimates a continuous value
- Classification predicts a discrete category

MNIST: classify hand-written digits
(10 classes)

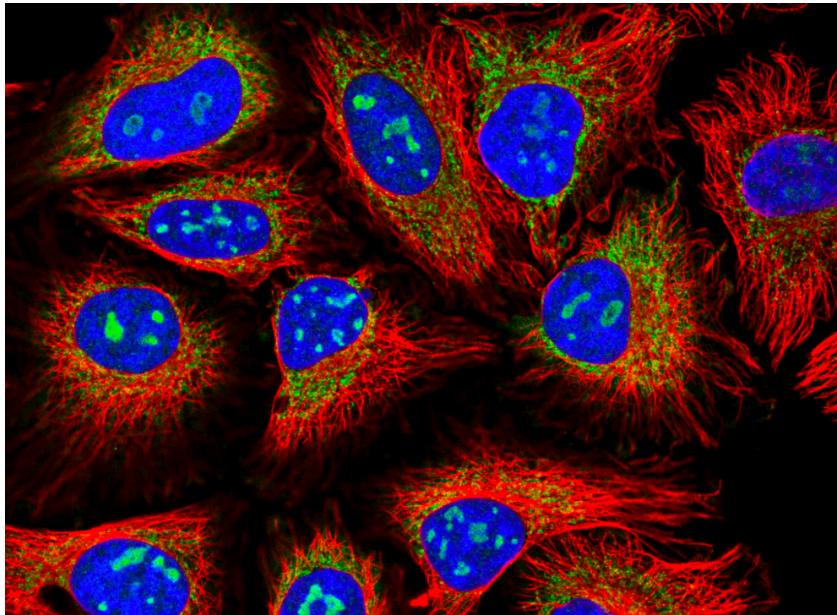


ImageNet: classify nature objects
(1000 classes)



Classification Tasks at Kaggle

Classify human protein microscope images into 28 categories



- 0. Nucleoplasm
- 1. Nuclear membrane
- 2. Nucleoli
- 3. Nucleoli fibrillar
- 4. Nuclear speckles
- 5. Nuclear bodies
- 6. Endoplasmic reticu
- 7. Golgi apparatus
- 8. Peroxisomes
- 9. Endosomes
- 10. Lysosomes
- 11. Intermediate fila
- 12. Actin filaments
- 13. Focal adhesion si
- 14. Microtubules
- 15. Microtubule ends
- 16. Cytokinetic bridg

<https://www.kaggle.com/c/human-protein-atlas-image-classification>

Classification Tasks at Kaggle

Classify malware into 9 categories



<https://www.kaggle.com/c/malware-classification>

Classification Tasks at Kaggle

Classify toxic Wikipedia comments into 7 categories

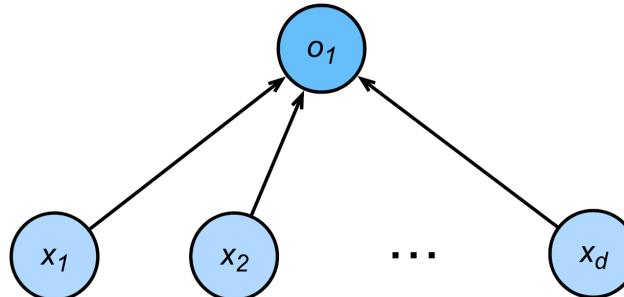
comment_text	toxic	severe_toxic	obscene
Explanation\nWhy the edits made under my user...	0	0	0
D'aww! He matches this background colour I'm s...	0	0	0
Hey man, I'm really not trying to edit war. It...	0	0	0
"\nMore\nI can't make any real suggestions on ...	0	0	0
You, sir, are my hero. Any chance you remember...	0	0	0

<https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>

From Regression to Multi-class Classification

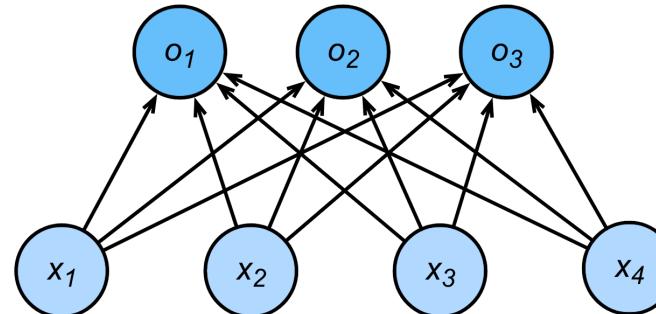
Regression

- Single continuous output
- Natural scale in \mathbb{R}
- Loss given e.g. in terms of difference $y - f(x)$



Classification

- Multiple classes, typically multiple outputs
- Score *should* reflect confidence ...



From Regression to Multi-class Classification

Square Loss

- One hot encoding per class

$$\mathbf{y} = [y_1, y_2, \dots, y_n]^\top$$

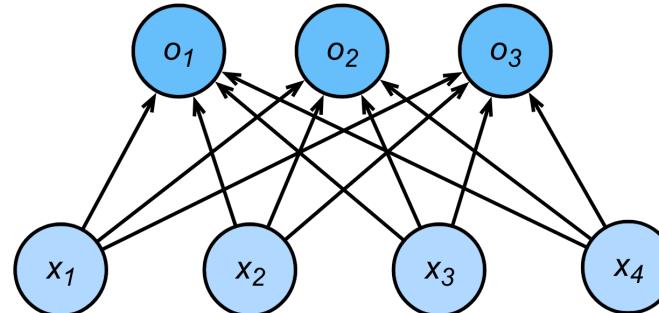
$$y_i = \begin{cases} 1 & \text{if } i = y \\ 0 & \text{otherwise} \end{cases}$$

- Train with squared loss
- Largest output wins

$$\hat{y} = \operatorname{argmax}_i o_i$$

Classification

- Multiple classes, typically multiple outputs
- Score *should* reflect confidence ...



From Regression to Multi-class Classification

Uncalibrated Scale

- One output per class
- Largest output wins

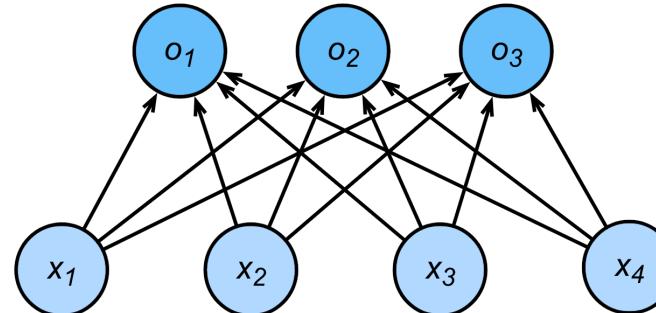
$$\hat{y} = \operatorname{argmax}_i o_i$$

- Want that correct class is recognized confidently
(large margin)

$$o_y - o_i \geq \Delta(y, i)$$

Classification

- Multiple classes, typically multiple outputs
- Score *should* reflect confidence ...



From Regression to Multi-class Classification

Calibrated Scale

- Output matches probabilities (nonnegative, sums to 1)

$$p(y|o) = \text{softmax}(o)$$

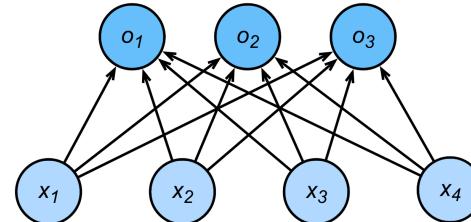
$$= \frac{\exp(o_y)}{\sum_i \exp(o_i)}$$

- Negative log-likelihood

$$-\log p(y|o) = \log \sum_i \exp(o_i) - o_y$$

Classification

- Multiple classes, typically multiple outputs
- Score *should* reflect confidence ...



Softmax and Cross-Entropy Loss

- Negative log-likelihood (for given label y)

$$-\log p(y|o) = \log \sum_i \exp(o_i) - o_y$$

- Cross-Entropy Loss (for probability distribution y)

$$l(y, o) = \log \sum_i \exp(o_i) - y^\top o$$

- Gradient

Difference between true
and estimated probability

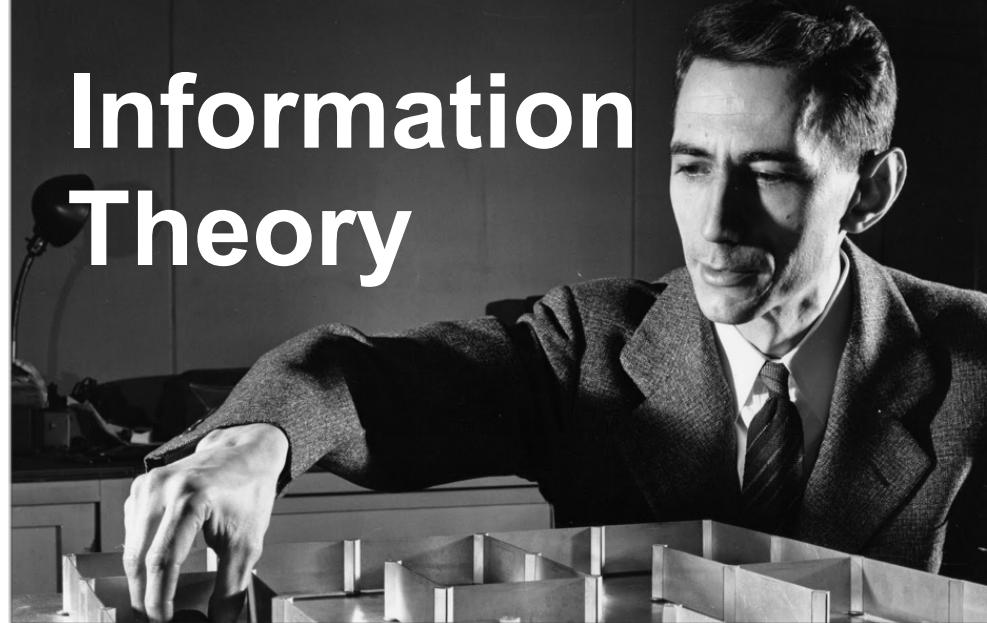
$$\partial_o l(y, o) = \frac{\exp(o)}{\sum_i \exp(o_i)} - y$$





MAGIC Etch A Sketch® SCREEN

Information Theory



Projection
Dial

OHIO ART

The World of Toys®

Vertical
Dial

MAGIC SCREEN IS GLASS SET IN STURDY PLASTIC FRAME
DO NOT USE WITH CARE

courses.

aws

Entropy

- Data source producing observations $x_1 \dots x_n$
- **How much ‘information’ is in this source?**
 - Tossing a fair coin - at each step the surprise is whether it’s heads or tails
 - Rolling a fair dice - we have 1 out of 6 outcomes. This should be *more* surprising than the dice
 - Picture of a white wall vs. picture of a football stadium (the football stadium should have more information)
- **Measure is minimum number of bits needed**

Entropy

- Data source producing data $x_1 \dots x_n$ with probability $p(x)$
- **Definition**

$$H[p] = - \sum_j p_j \log p_j$$

- **Coding theorem**

Entropy is lower bound on bits (or rather nats - base e)

$$2^a = e^b \text{ hence } a \log 2 = b \text{ hence bits} = \frac{H[p]}{\log 2}$$

- Entropy is concave since $p \log p$ is convex

Entropy (binary form)

- Fair coin ($p = 0.5$)

$$H[p] = -0.5 \cdot \log_2 0.5 - 0.5 \cdot \log_2 0.5 = 1 \text{ bit}$$

- Biased coin ($p = 0.9$)

$$H[p] = -0.9 \cdot \log_2 0.9 - 0.1 \cdot \log_2 0.1 = 0.47 \text{ bit}$$

- Dungeons and Dragons (20-sided dice)

$$H[p] = -\log_2 \frac{1}{20} = 4.32 \text{ bit}$$



Kraft Inequality

- **Prefix Code**

- Map x to code $c(x)$ with length $l(x)$
- No $c(x)$ is the prefix for any $c(x')$,
e.g. STAT and STATISTICS cannot both be codewords

$$\begin{cases} a \rightarrow 0 \\ b \rightarrow 01 \\ c \rightarrow 011 \\ d \rightarrow 0111 \end{cases}$$

$$\begin{cases} a \rightarrow 0 \\ b \rightarrow 10 \\ c \rightarrow 110 \\ d \rightarrow 111 \end{cases}$$

Kraft Inequality

- **Inequality**

$$1 \geq \sum_x 2^{-l(x)} \text{ if and only if prefix code}$$

- **Proof**

Generate random string. Probability that it's a codeword

$$1 \geq \Pr(\text{collision}) = \sum_x \Pr(x \text{ is a hit}) = \sum_x 2^{-l(x)}$$

For converse explicitly construct prefix code recursively

- Pick set of $\{x\}$ with smallest $l(x)$ and generate code
- Use leftovers and break them up into sets of weight $2^{-l(x)}$
- Give each of them prefix and rescale by $2^{l(x)}$

Coding Theorem Proof

- Entropy is lower bound on number of bits

$$\text{bits}(p) = \frac{H[p]}{\log 2} = - \sum_j p_j \log_2 p_j$$

- Generate prefix code with length $l(x) = \lceil \log_2 p(x) \rceil$
This is within 1 bit of optimal code
Kraft inequality shows that such a thing exists.
- Combine data in k-tuples and encode
This is within $1/k$ bit of optimal code.
- Optimality proof via KL divergence (omitted here)

Kullback-Leibler Divergence

- Distance between distributions (e.g. truth & estimate)

Number of extra bits when using the wrong code

$$D[p\|q] = \int dp(x) \log \frac{p(x)}{q(x)} = \int dp(x) [-\log q(x)] - [-\log p(x)]$$

Inefficient bits

Optimal bits

- Nonnegativity of KL Divergence

$$D[p\|p] = \int dp(x) \log \frac{p(x)}{p(x)} = 0$$

Jensen Inequality

$$D[p\|q] = - \int dp(x) \log \frac{q(x)}{p(x)} \geq - \log \int dp(x) \frac{q(x)}{p(x)} = 0$$

Back to the Cross Entropy Loss

- Cross entropy loss

$$l(y, o) = \log \sum_i \exp(o_i) - y^\top o$$

- Kullback Leiber divergence

$$\begin{aligned} D(\text{softmax}(o) \| q) &= \sum_i q_i \log q_i - q_i \log \text{softmax}(o)_i \\ &= -H[q] + \log \sum_i \exp(o_i) - \sum_i q_i o_i \end{aligned}$$

Independent of o

Extended Reading

- Cover and Thomas (Elements of Information Theory)
dl.acm.org/citation.cfm?id=1146355
- Information theory course (Entropy primer)
spl.cse.nsysu.edu.tw/cpchen/courses/ita/l1_entropy.pdf
- David MacKay (Information Theory and Learning)
www.inference.org.uk/iitprnn/book.html
- Conditional Entropy, Mutual Information,
Exponential Families, Maximum Entropy Estimation

Summary

- **Maximum Likelihood**
 - Gauss and means
 - More loss functions (l_1 loss, trimmed mean)
 - Regression revisited
- **Classification**
 - Computing discrete probabilities
 - Likelihood and loss functions
- **Information Theory**