CSC311: Midterm Review

Julyan Keller-Baruch

February 13 2020

Based on slides from Anastasia Razdaibiedina, Sargur Srihari, James Lucas and others

Midterm review

For midterm review we'll go through:

1. Important ML concepts
2. Exercises

ML concepts

- **What is supervised learning?**

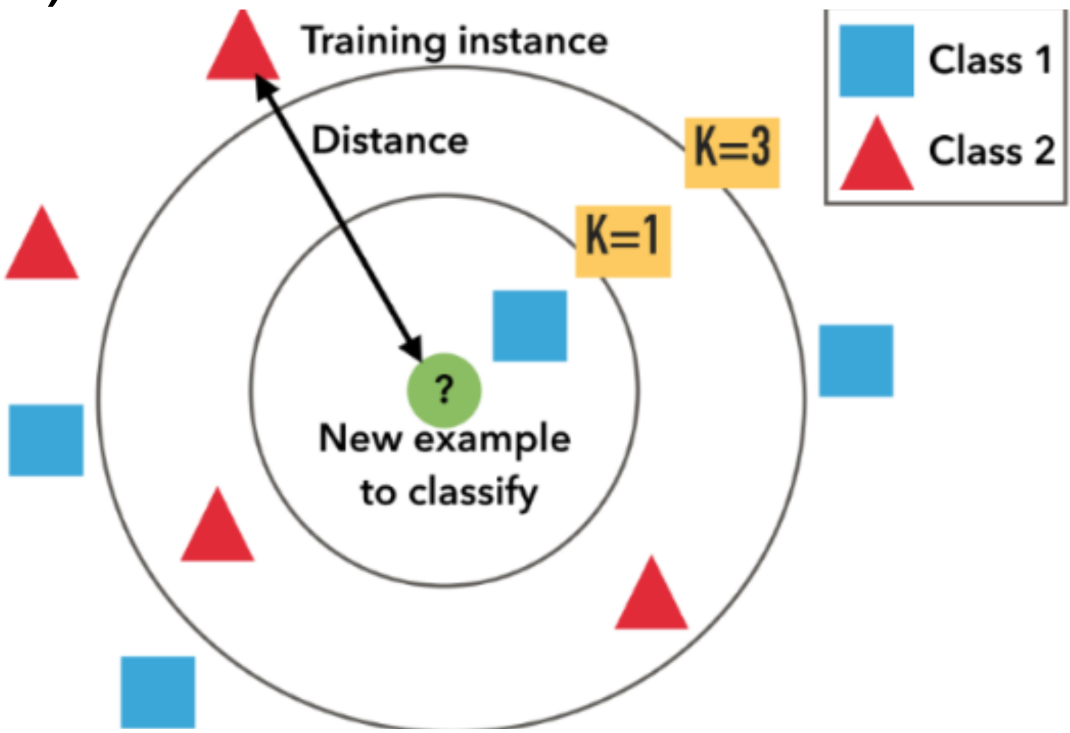
Answer: ML setting when our training set consists of inputs and their corresponding labels.

- **Difference between regression / classification?**

Answer: in ***classification*** we are predicting a discrete target (like cat or dog class), while in ***regression*** we are predicting a continuous-valued target (like temperature).

- **What does kNN do?**

Answer: k Nearest Neighbours is an algorithm that predicts value of a new example based on its k nearest labeled neighbours.



ML concepts

- **How does decision tree work?**

Answer: decision trees make predictions by sequentially splitting data on different attributes.

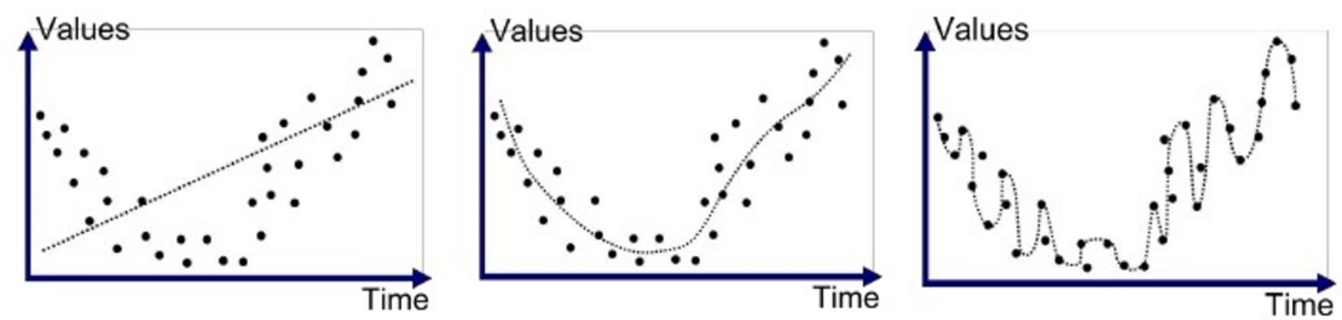
- **Name 2 advantages of kNN vs decision tree and vice versa.**

kNN: can incorporate interesting distance measures; few hyperparameters

decision trees: fast at test time; more interpretable; better deals with missing values.

- **What is overfitting and underfitting?**

Overfitting: When the model gets good performance on a particular dataset by "memorizing" it, but fails to generalize to new data.



Underfitted   Good Fit/Robust   Overfitted

- **Why do we need a validation set?**

Answer: to prevent overfitting.

ML concepts

- **Based on which measure we can choose a good decision tree split?**

Answer:

Fitting the tree is finding an order to split the data, such that the information gain is maximized at each split.

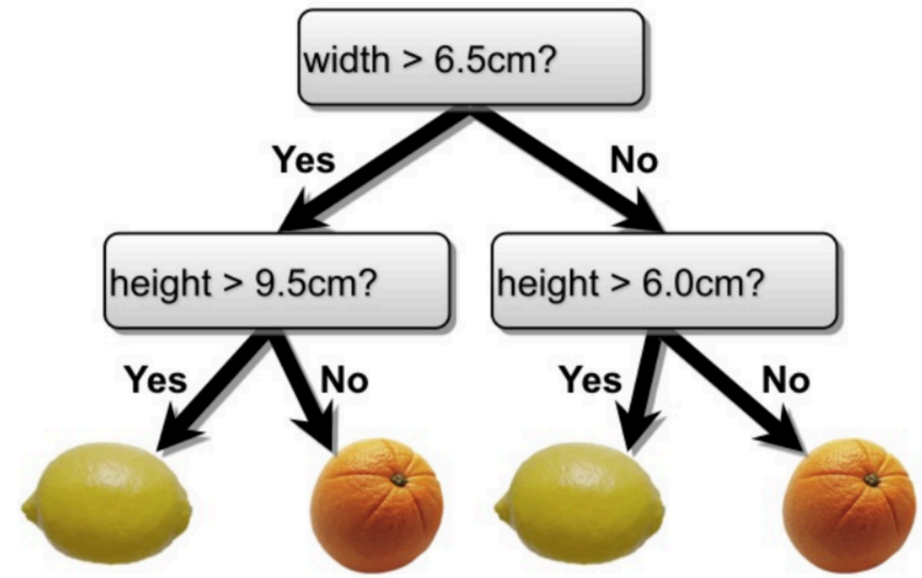


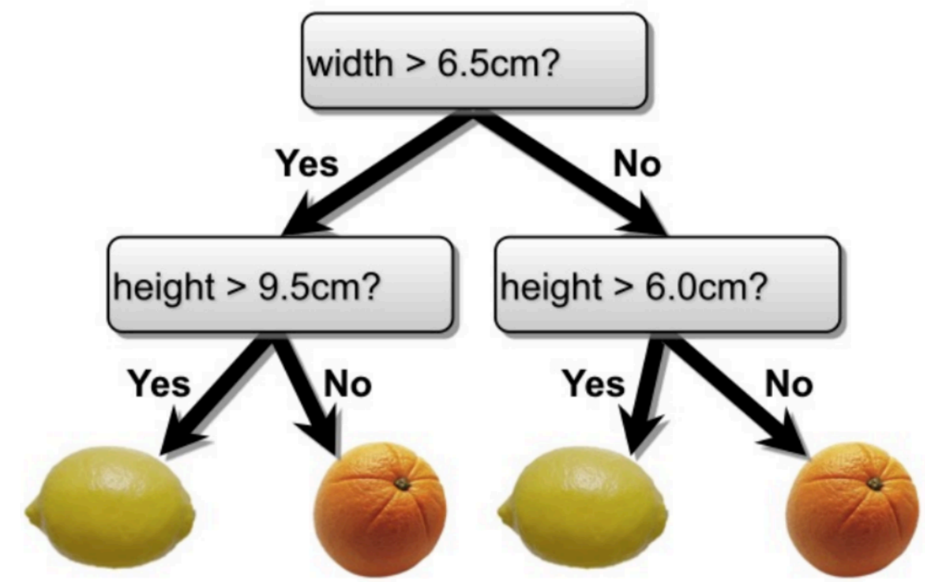
Image source: http://www.cs.toronto.edu/~rgrosse/courses/csc2515_2019/tutorials/tut7/Midterm_Review_Tutorial.pdf

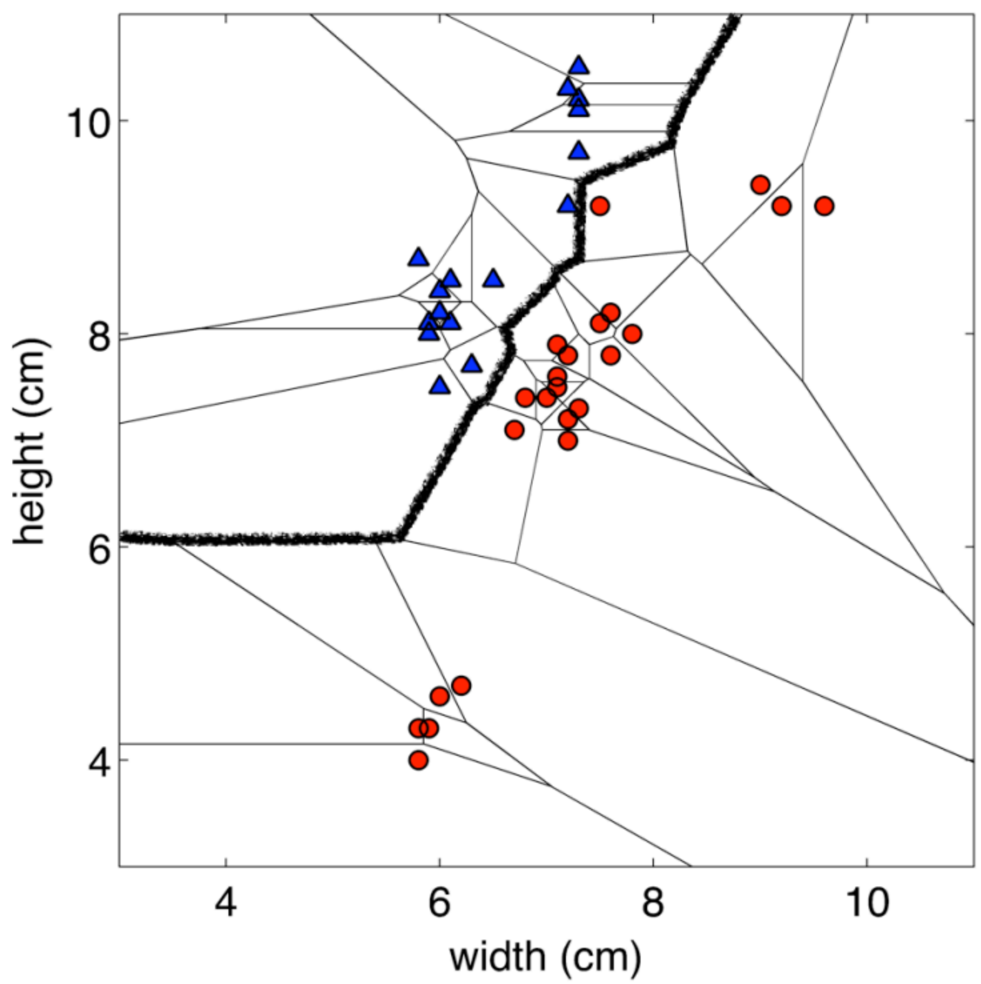
Information gain: tells us how much "information" a feature gives us about the class.

Entropy: a measure of impurity, disorder or uncertainty in a set of examples. "How unpredictable a dataset is".

ML concepts

- **Decision boundary of decision trees vs. kNN?**





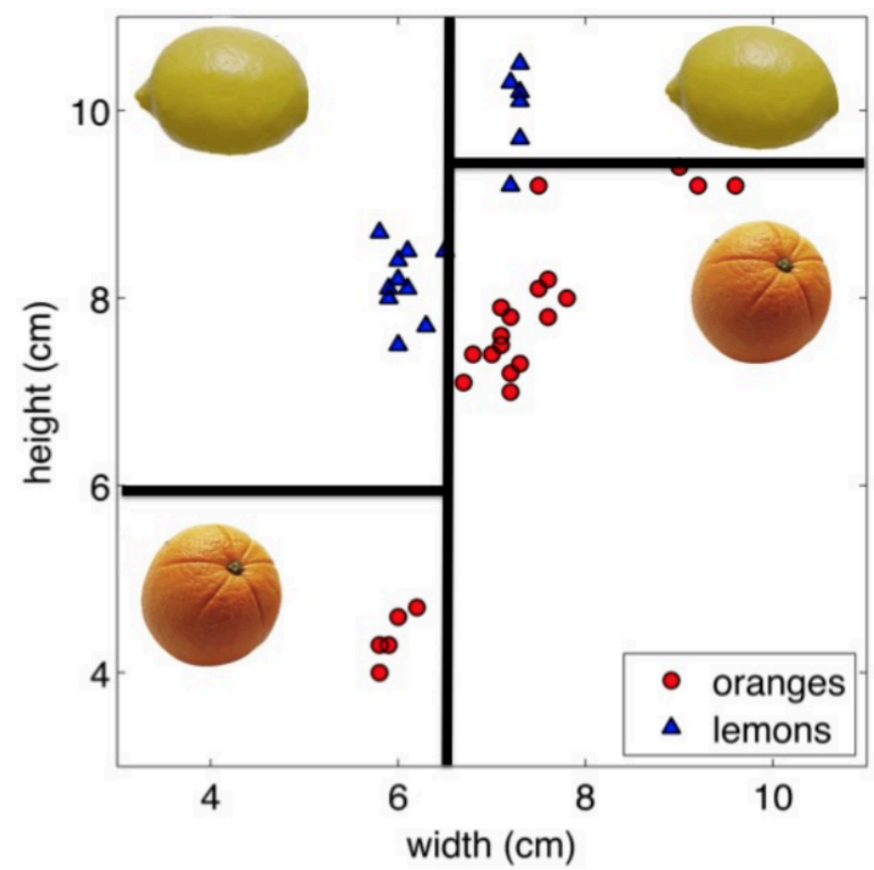
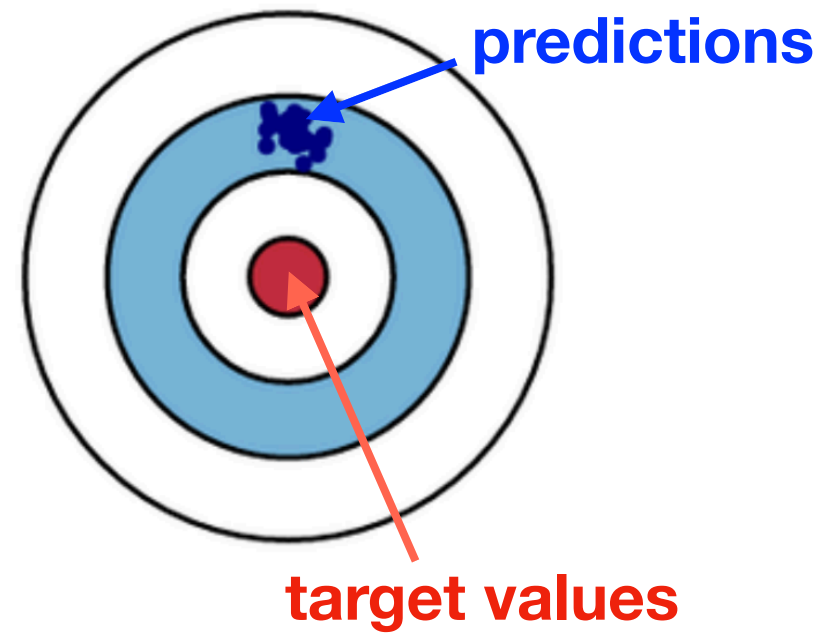


Image source:
http://www.cs.toronto.edu/~rgrosse/courses/csc2515_2019/tutorials/tut7/Midterm_Review_Tutorial.pdf

ML concepts

- **What does this picture tells us about our data (bias / variance)?**



Answer: high bias & low variance.

ML concepts

- **Are decision trees and kNN supervised / unsupervised algorithms?**

Answer: supervised (we need labels).

ML concepts

- **Write a model for binary linear classification …**

$$z = \mathbf{w}^T \mathbf{x} + b$$

$$y = \begin{cases} 1 & \text{if } z \geq r \\ 0 & \text{if } z < r \end{cases}$$

- **What are the two ways of finding good values for model's parameters (w, b)?**

  A. direct solution

  B. iterative solution (gradient descent)

- **What is loss function?**

Answer: it's a function that evaluates how well specific algorithm models the given data; loss function takes predicted values and target values as inputs.

ML concepts

- **A loss function for linear classification:** 0-1 loss

- **Problem?**

<u>Answer</u>: 0-1 loss is bad because it's not informative - its derivative is 0 everywhere it's defined.

$$\mathcal{L}_{0-1}(y,t) = \begin{cases} 0 & \text{if } y = t \\ 1 & \text{if } y \neq t \end{cases}$$

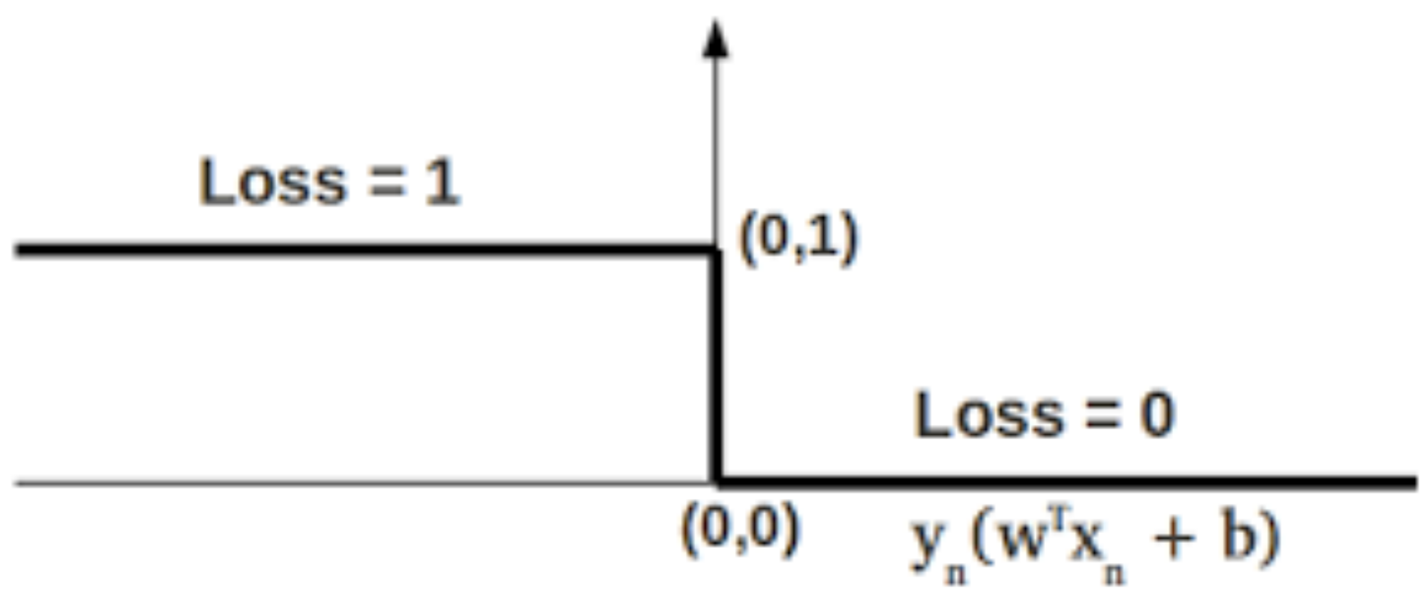
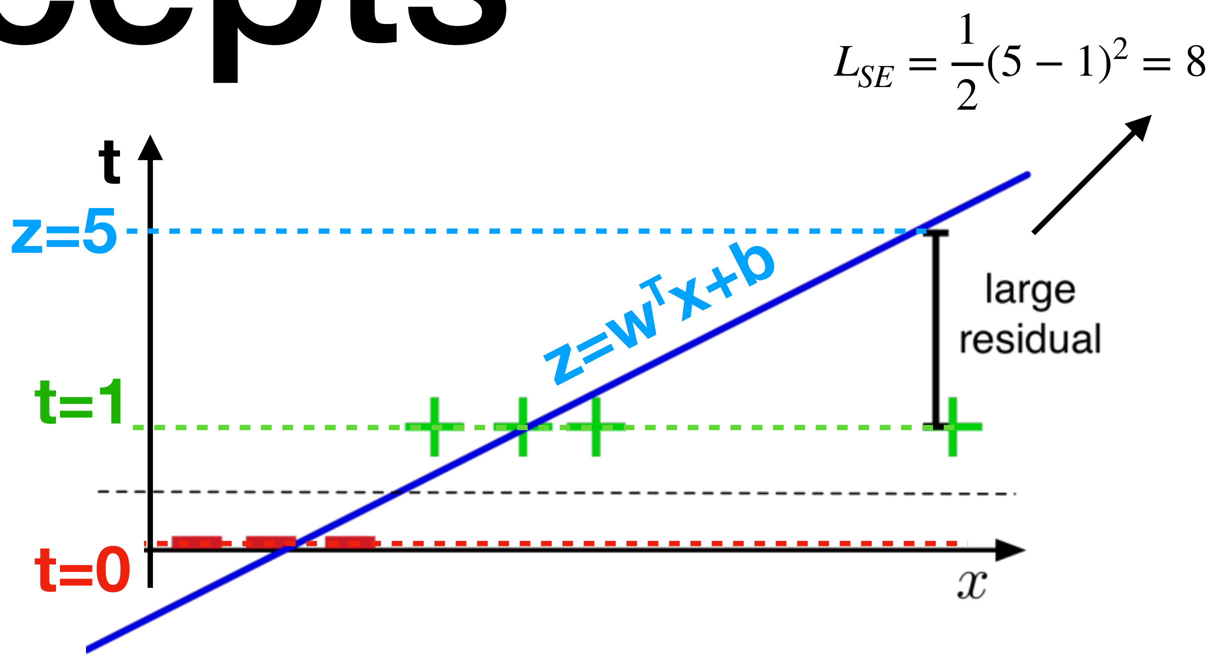


Image source: http://www.cs.umd.edu/class/spring2017/cmsc422/slides0101/lecture11.pdf

ML concepts

- **What are the problems with squared error loss function in classification?**

<u>Answer</u>: squared error loss gives a big penalty for correct predictions that are made with high confidence.



A solution?

Predict values only in [0,1] interval. For that we use **sigmoid function** to squash y into [0,1]:

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

$$z = \mathbf{w}^\top \mathbf{x} + b$$

$$y = \sigma(z)$$

$$\mathcal{L}_{\mathrm{SE}}(y,t) = \frac{1}{2}(y-t)^2.$$

(t=1)

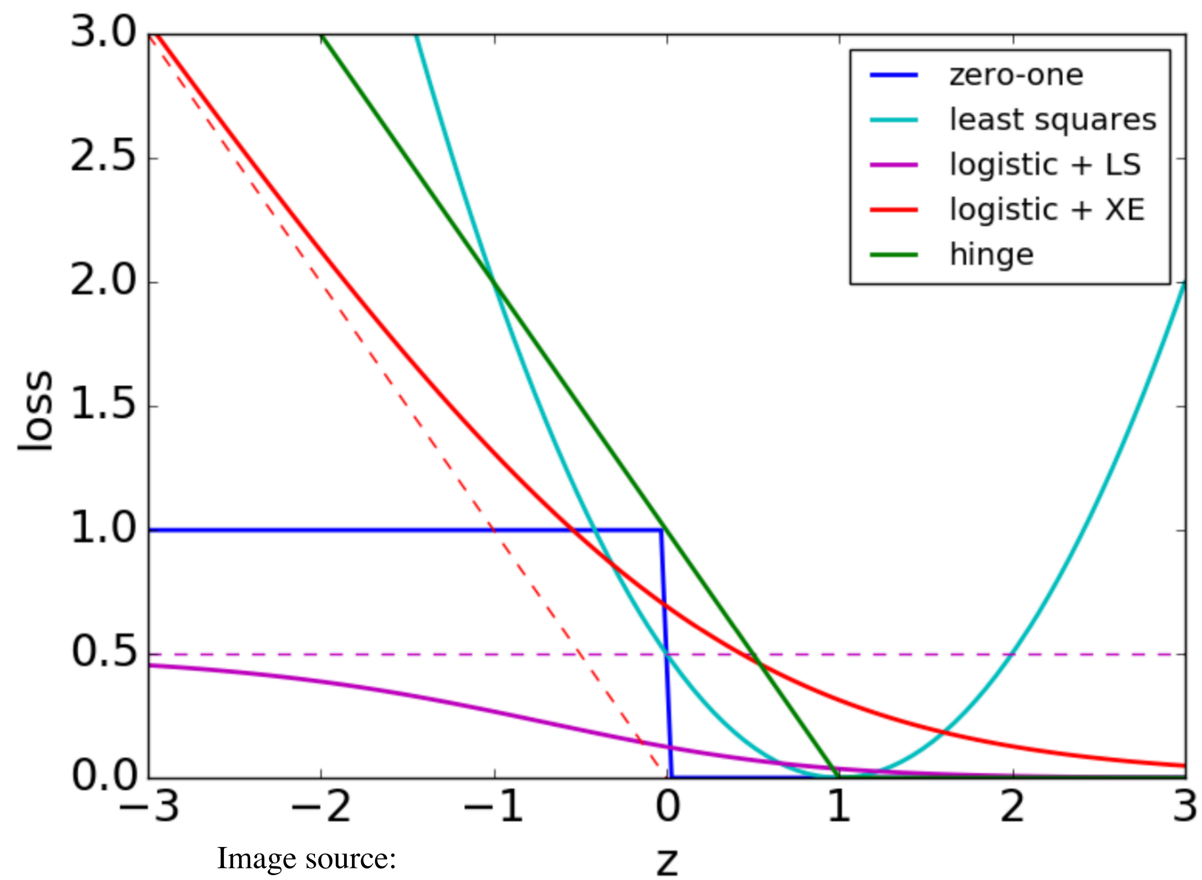


Image source:
http://www.cs.toronto.edu/~rgrosse/courses/csc321_2017/slides/lec4.pdf

ML concepts

Another solution:

Cross entropy loss.

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

$$z = \mathbf{w}^\top \mathbf{x} + b$$
$$y = \sigma(z)$$

$$L_{CE} = -t\log(y) - (1-t)\log(1-y)$$

(t=1)

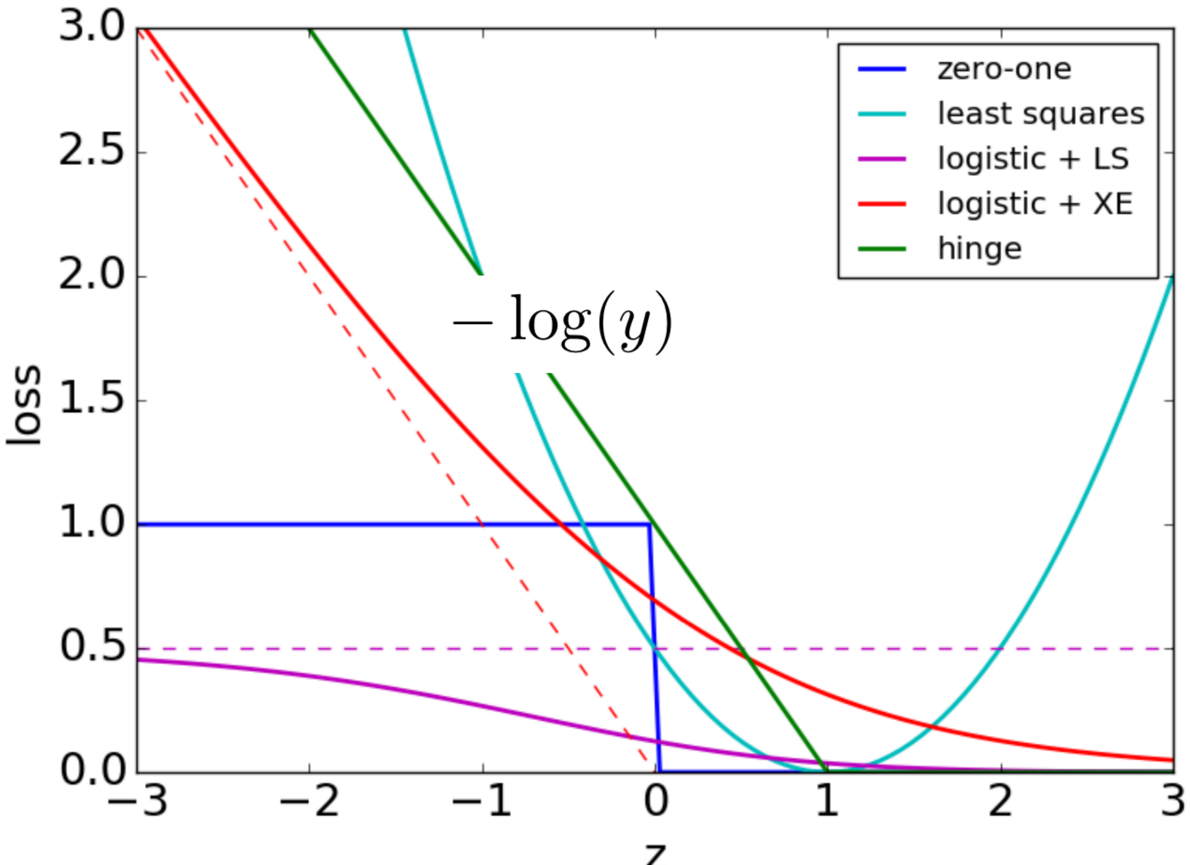


Image source:
http://www.cs.toronto.edu/~rgrosse/courses/csc321_2017/slides/lec4.pdf

ML concepts

- **What is the difference between parameters / hyper parameters of the model?**

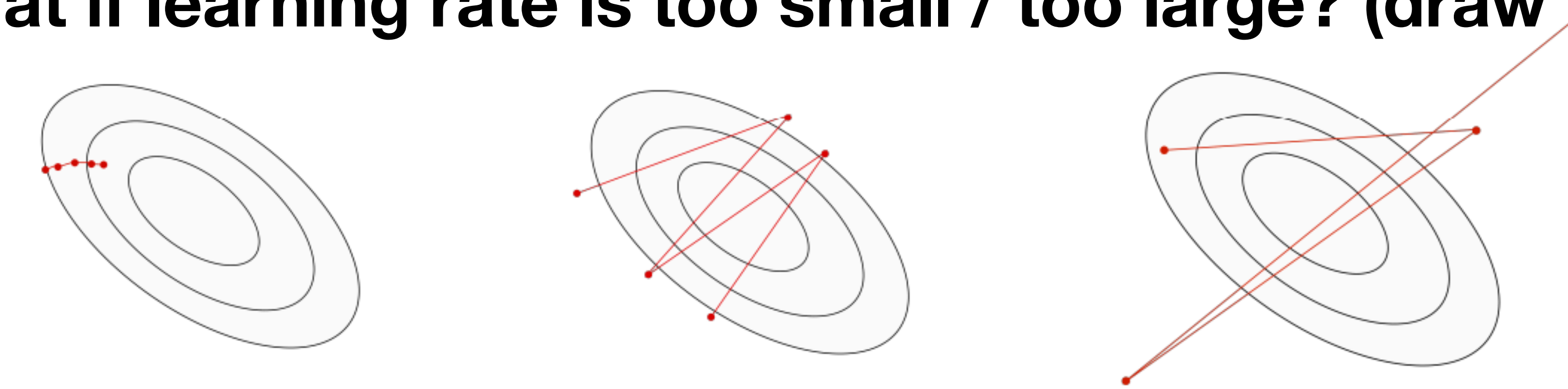
Answer: ***parameters*** are learned through training (by iteratively performing gradient descent updates) - weights and biases, ***hyperparameters*** are "manually" adjusted and set before training - number of hidden layers of a neural network, k for kNN, learning rate etc.

- **What is learning rate?**

Answer: learning rate is a hyper parameter that controls how much the weights are updated at each iteration.

$$w_j \leftarrow w_j - \alpha \frac{\partial \mathcal{J}}{\partial w_j}$$

- **What if learning rate is too small / too large? (draw a picture)**



α too small: slow progress

α too large: oscillations

α much too large: instability

ML concepts

- **What is regularization? Why do we need it?**

Answer: regularization is a technique of adding an extra term to the loss function. It reduces overfitting by keeping the weights of the model smaller.

L1 vs L2 Regularization

L1:

$$O = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} X_{i,j} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

L2:

$$O = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} X_{i,j} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^{\,2}$$

ML concepts

- **What is softmax? Calculate** $softmax(\begin{bmatrix} 2 \\ 1 \\ 0.1 \end{bmatrix})$

<u>Answer</u>: softmax is an ***activation function*** for multi-class classification that maps input ***logits*** to probabilities.

$$softmax(\begin{bmatrix} 2 \\ 1 \\ 0.1 \end{bmatrix}) = \begin{bmatrix} e^2/(e^2 + e^1 + e^{0.1}) \\ e^1/(e^2 + e^1 + e^{0.1}) \\ e^{0.1}/(e^2 + e^1 + e^{0.1}) \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix}$$

$$y_k = \mathrm{softmax}(z_1, \ldots, z_K)_k = \frac{e^{z_k}}{\sum_{k'} e^{z_{k'}}}$$

Other topics to know

- Difference between training, validation and testing sets
- Maximum Likelihood estimation (Slides 3.26-3.30)
- Bagging
- Responsible for material up until slide 5.11

Example 1 - linear classifier weights

Find a linear classifier with weights w1, w2, w3, and b which correctly classifies all of these training examples:

x_1	x_2	x_3	t
0	0	0	1
0	1	0	0
0	1	1	1
1	1	1	0

$$w_1 x_1 + w_2 x_2 + w_3 x_3 + b \gtrless 0$$

<u>Answer</u>: write a system of inequalities and find one solution (there would be many possible answers).

$$\begin{aligned} b &> 0 \\ w_2 + b &< 0 \\ w_2 + w_3 + b &> 0 \\ w_1 + w_2 + w_3 + b &< 0 \end{aligned} \qquad \begin{aligned} b &= 1 \\ w_1 &= -2 \\ w_2 &= -2 \\ w_3 &= 2 \end{aligned}$$

Example 2 - entropy

Suppose binary-valued random variables X and Y have the following joint distribution:

	$Y=0$	$Y=1$
$X=0$	$1/8$	$3/8$
$X=1$	$2/8$	$2/8$

Find ***entropy*** of a joint distribution $H(X,Y)$ and ***conditional entropy*** of Y given X=0.

<u>Answer</u>:

entropy of a joint distribution $H(X,Y) = -\sum_x \sum_y p(X=x,Y=y)\log_2 p(X=x,Y=y)$

$$H(X,Y) = -\frac{1}{8}\log_2\frac{1}{8} - \frac{3}{8}\log_2\frac{3}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{2}{8}\log_2\frac{2}{8}$$

conditional entropy $H(Y|X=0) = -\sum_{y\in\{0,1\}} p(Y=y|X=0)\cdot\log_2 p(Y=y|X=0)$

	$Y=0$	$Y=1$
$X=0$	$1/8$	$3/8$
$X=1$	$2/8$	$2/8$

	$Y=0$	$Y=1$
$X=0$	1/4	3/4

$$p(Y=0|X=0) = \frac{p(Y=0,X=0)}{p(X=0)} = \frac{1/8}{4/8} = \frac{1}{4}$$

$$H(Y|X=0) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$$

Example 3 - Information Gain

Suppose binary-valued random variables X and Y have the following joint distribution:

	$Y=0$	$Y=1$
$X=0$	$1/8$	$3/8$
$X=1$	$2/8$	$2/8$

Find ***information gain*** $IG(Y|X)$.

<u>Answer</u>: **Information Gain:** $IG(Y|X) = H(Y) - H(Y|X)$

	$Y=0$	$Y=1$
$X=0$	$1/8$	$3/8$
$X=1$	$2/8$	$2/8$
	3/8	**5/8**

p(Y=0)

$$H(Y) = -\sum_y p(Y=y)\log_2 p(Y=y)$$

$$H(Y) = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8}$$

	$Y=0$	$Y=1$	
$X=0$	$1/8$	$3/8$	**4/8**
$X=1$	$2/8$	$2/8$	**4/8**

$$H(Y|X) = p(X=0)\cdot H(Y|X=0) + p(X=1)\cdot H(Y|X=1)$$

$$p(X=0) = \frac{4}{8} = \frac{1}{2} \qquad p(X=1) = \frac{4}{8} = \frac{1}{2}$$

$$H(Y|X=0) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$$

$$H(Y|X=1) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$$

plug in values into H(Y|X) equation