2021 Fast.ai Community Course

QUEENSLAND HUB

QUEENSLAND AI

Lesson 5 – Data Ethics and Other DL concepts

Notebooks: 05_pet_breeds.ipynb
08_collab.ipynb

Key concepts

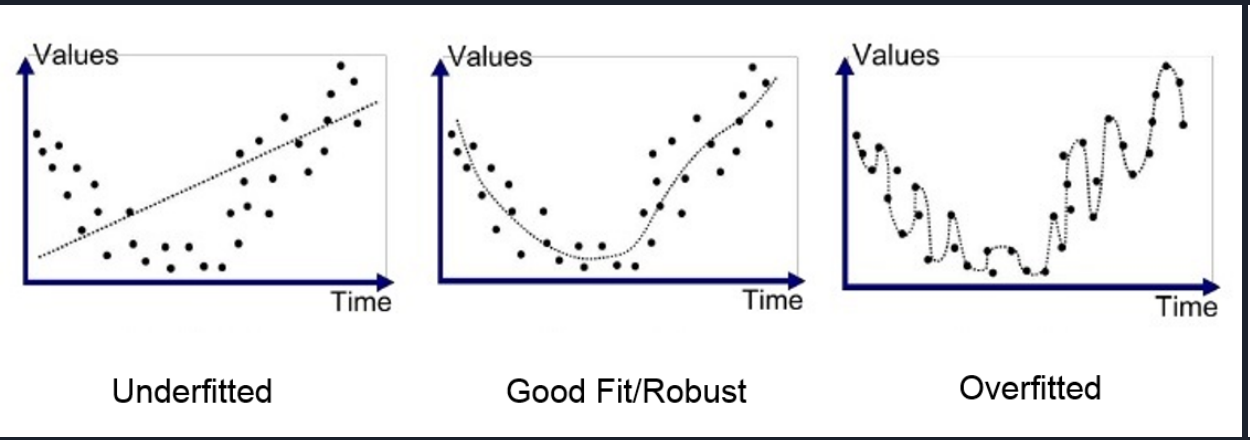
- Regularisation by weight decay
- Data augmentation
- Softmax
- Cross entropy loss
- Class feedback survey

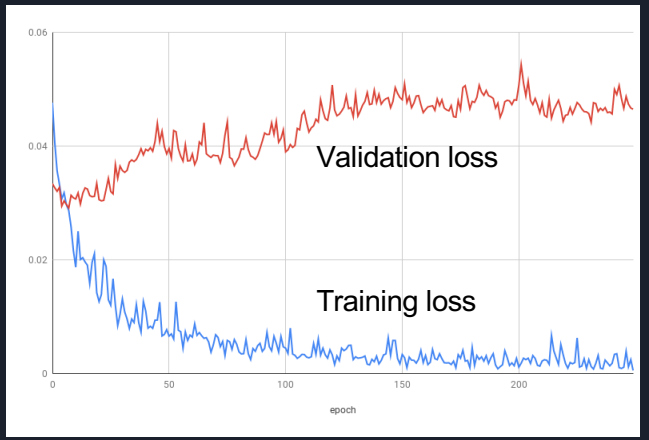
Capacity of a neural net

- If we train a resnet18 model (11 million weights) on the MNIST data for 10 epochs, what can go wrong?

Overfitting

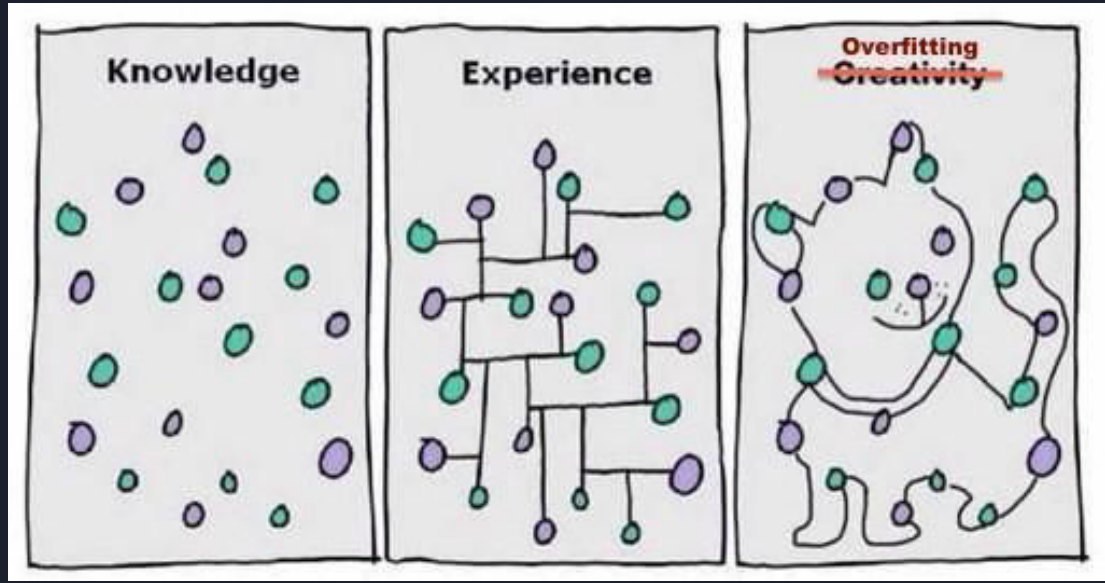
- What is overfitting?
- Deep neural networks can be prone to overfitting, why?
- What are possible ways to address overfitting?





Overfitting vs underfitting

- Underfitting is like horoscopes, making such vague, flexible predictions that sound vaguely true but have low predictive power.
- Overfitting is like Ptolemaic astronomy (the geocentric model based on a complicated system of nested circular orbits), which required more and more circles to explain deviation.
- Can you come up with other analogies? (discuss at end of class)



Regularisation by weight decay

- Constrain the complexity of the model by ensuring the weights of the model remain small.
- Penalize the model during training based on the magnitude of the weights.

$$L = L_0 + \alpha \sum_w w^2$$

α is the regularization parameter

```
learn.fit_one_cycle(5, 5e-3, wd=0.1)
```

Probabilistic interpretation of weight decay

- Recall from last lesson the cross-entropy loss is the negative log likehood of getting the correct class P(D|w,x)

$$P(D|w,x) = \exp(-L_0)$$

- Similarly the regulariser can be interpreted in terms of the log of a Gaussian prior probability distribution over w:

$$P(w) \propto \exp(-\alpha w^2)$$

- The new loss function $L = L_0 + \alpha \Sigma w^2$ corresponds to the log of the posterior probability of w:

$$P(w|D) \propto P(D|w) * P(w) = \exp(-L)$$

Bayes' Theorem!

- w* found by minimizing L can be interpreted as the *maximum a posteriori* (MAP) estimate of w (as opposed to maximum likelihood estimate if using L_0)

Data augmentation

- List of possible augmentations https://docs.fast.ai/vision.augment.html

```
pets = DataBlock(blocks = (ImageBlock, CategoryBlock),
                 get_items=get_image_files,
                 splitter=RandomSplitter(seed=42),
                 get_y=using_attr(RegexLabeller(r'(.+)_\d+.jpg$'), 'name'),
                 item_tfms=Resize(460),
                 batch_tfms=aug_transforms(size=224, min_scale=0.75))
dls = pets.dataloaders(path/"images")
```

```
batch_tfms=aug_transforms(mult=1.0, do_flip=True,
                          flip_vert=True, max_rotate=10.,
                          max_zoom=1.1,max_warp=0.2,
                          p_affine=0.75, p_lighting=0,
                          xtra_tfms=None))
```

Caveat

- How can data augmentation go wrong?

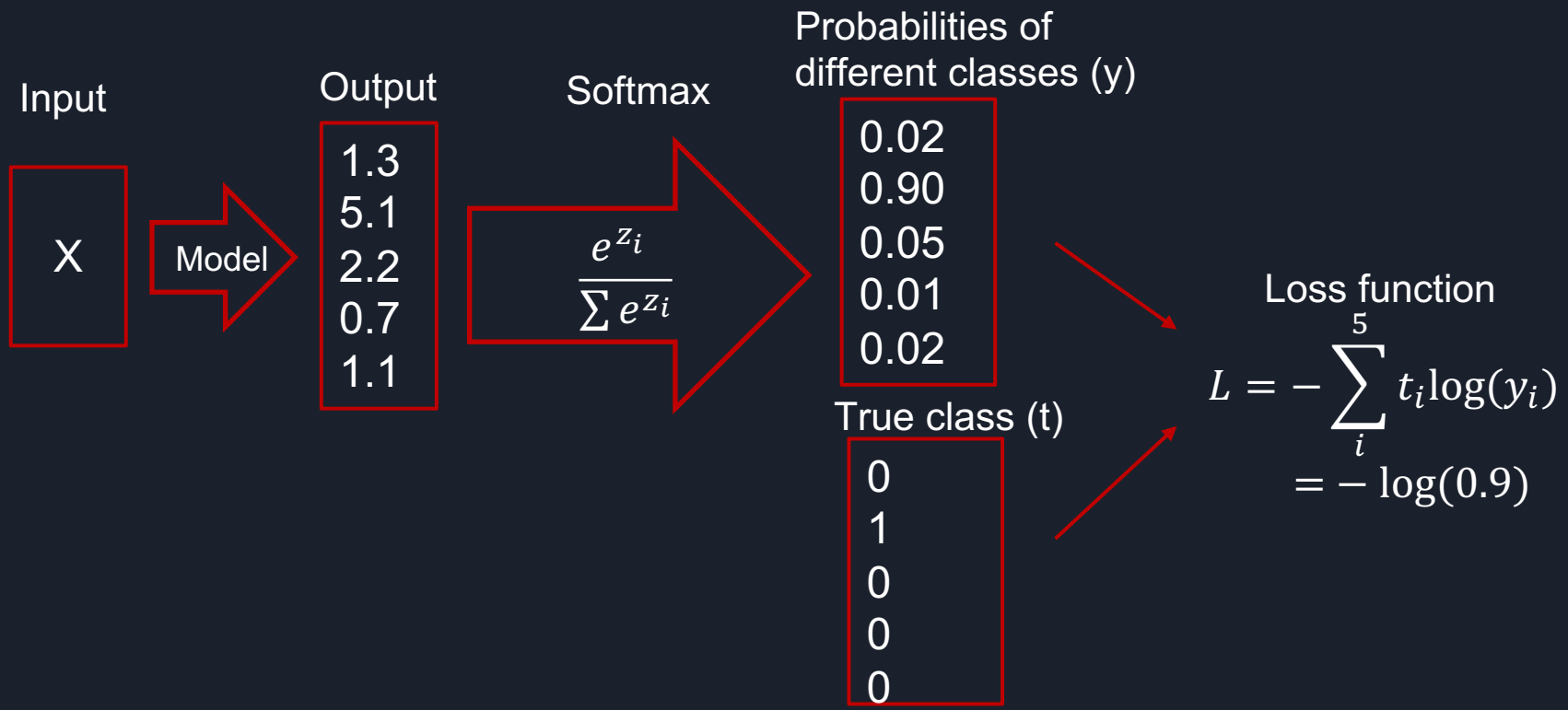
Other methods to address overfitting

- L1 Regularization
- Dropout
- Early Stopping

(discuss with mentors after class if interested)

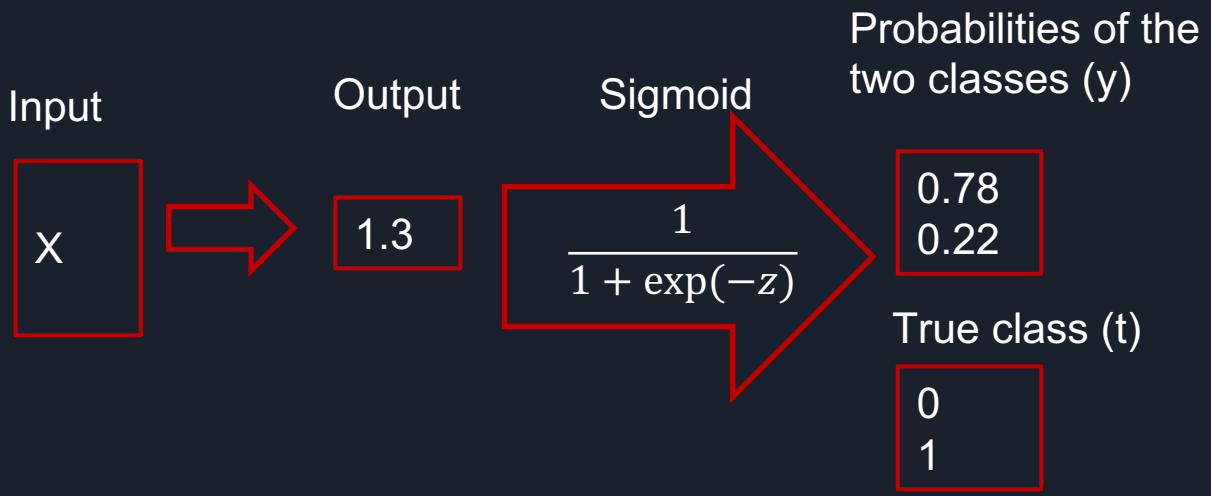
Softmax

- The softmax function turns the raw outputs into the probabilities of different classes



Binary cross entropy

- Binary Cross Entropy loss is only a special case of cross entropy for binary classification.



$$L = -\sum_i^2 t_i \log(y_i) = -[t \log(y) + (1-t)(1-\log(y))] = -\log(0.22)$$

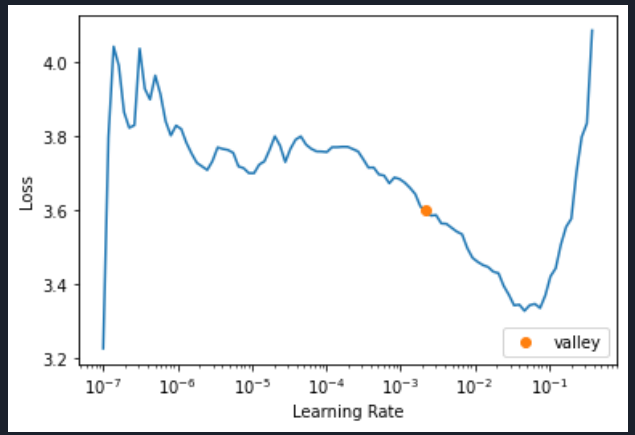
Exercise 1

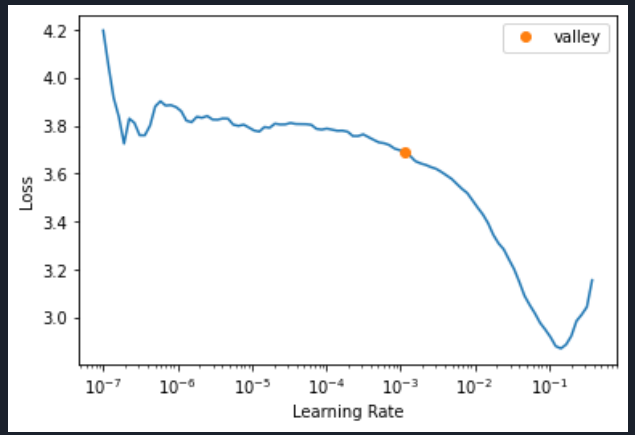
- Team 1: pen & paper vs Team 2: find the solution through code

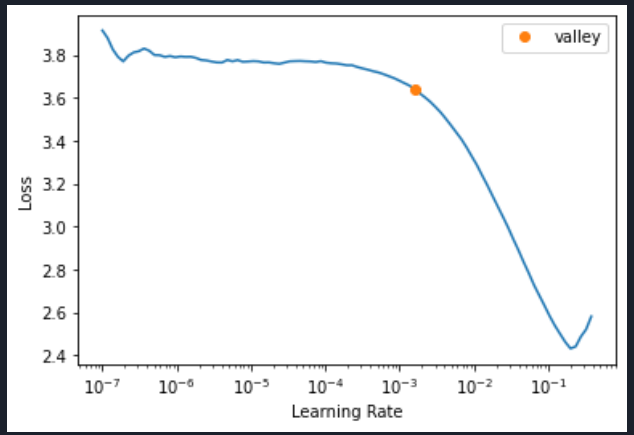
- Recall the simple net from last week on MNIST: one linear layer, RELU and another linear layer.
- Write the forward pass, prediction and loss function.
- Given your first layer has 30 nodes, how many weights does your model have?
- If the batch size is 100, what is the size of your output?
- What might be the disadvantage of this model architecture?

Exercise 2

- Team 3: pen & paper vs Team 4: find the solution through code
- These 3 plots are generated by learn.lr_find() on a resnet18 model trained on MNIST using three different batch sizes: 16, 64 and 512. Can you tell which plot corresponds to which batch size? Explain.
- Is a higher batch size a generally good thing? Think of the loss function landscape discussed last week.







Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour

[1606.02228] Systematic evaluation of CNN advances on the ImageNet

Exercise 3

- Team 5: pen & paper vs Team 6: find the solution through code

  - What happens when the weight decay parameter is very high?
  - When should weight decay not be used?

Class survey

- We would like to hear feedback from you on the course and mentors, please go to the poll here https://www.menti.com/gvsrpwspvc