Congratulations! You passed!

Grade received 100% **To pass** 80% or higher

Go to next item

Knowledge Distillation

Total points 7

1. The goal of knowledge distillation is optimizing the network implementation: **1 / 1 point**

- (•) False
- ( ) True

Correct
Exactly! Rather than optimizing, distillation seeks to create a more efficient model.

2. In knowledge distillation, the teacher will be trained using a _____. **1 / 1 point**

- ( ) GoogLeNet
- ( ) A Soft Target
- ( ) K-L divergence
- (•) A Standard objective function

Correct
Nailed it! This seeks to maximize the accuracy of the model.

3. DistilBERT is a bigger version of BERT with a modified architecture, but the same number of layers. **1 / 1 point**

- (•) No
- ( ) Yes

Correct
You're right! it's a smaller version of BERT: they reduced the numbers of layers and kept the rest of the architecture identical

4. In knowledge distillation, the "teacher" network is deployed in production as it is able to mimic the complex feature relationships of the "student" network. **1 / 1 point**

- ( ) True
- (•) False

Correct
Exactly! It's actually the "student" network the one deployed to mimic the "teacher" network.

5. For a multi-class classification problem, which ones of the following statements are true regarding the training cost functions of the "student" and the "teacher" networks? (Select all that apply) **1 / 1 point**

- [x] Soft targets encode more information about the knowledge learned by the teacher than its output class prediction per example.

Correct
That's right! Soft targets provide more information that the output class predicted per example as they include information about all the classes per training example through the probability distribution.

- [x] The teacher network is trained to maximize its accuracy and the the student network uses a cost function to approximate the probability distributions of the predictions of the teacher network.

Correct
That's right!

- [ ] They both share the same cost functions,
- [ ] The teacher network is trained to maximize its accuracy and the the student network uses a cost function to output the same classes as the teacher network.

6. When the softmax temperature ____, the soft targets defined by the teacher network become less informative **1 / 1 point**

- ( ) is equal to 1
- ( ) increases
- (•) decreases

Correct
That's right! The softness of the teacher's distribution is worse, thus less informative.

7. Generally, knowledge distillation is done by blending two loss functions and involves several hyperparameters. Here, L_h is the cross-entropy loss from the hard labels and LKL is the Kullback–Leibler divergence loss from the teacher labels. Which of the following statements are correct about the hyperparameters of knowledge distillation? (Select all that apply)

1 / 1 point

- [ ] In case of heavy data augmentation after training the teacher network, the alpha hyperparameter should be low in the student network loss function
- [x] When computing the the "standard" loss between the student's predicted class probabilities and the ground-truth "hard" labels, we use a value of the softmax temperature T equal to 1

  Correct
  That's right! This way, the student loss function would be a classical softmax function

- [x] In case of heavy data augmentation after training the teacher network, the alpha hyperparameter should be high in the student network loss function

  Correct
  That's correct! This high alpha parameter would reduce the influence of the hard labels that went through aggressive perturbations due to data augmentation

- [ ] When computing the the "standard" loss between the student's predicted class probabilities and the ground-truth "hard" labels, we use the same value of the softmax temperature T to compute the softmax on the teacher's logits