INFR10086 Machine Learning (MLG) Semester 1, 2022/23

Practice Exam

1. Discuss whether the following statements are true or false.

    a) If learning hypothesis class A has a larger sample complexity than learning hypothesis class B, then it requires more samples to find a model in A to achieve the same generalization error as finding a model in B.

    [*6 marks*]

    b) If hypothesis class A has a larger VC dimension than hypothesis class B, then the difference in training and test errors for models in class A is larger than those in class B.

    [*6 marks*]

    c) If model A has a lower test error than model B, then model A has a lower generalization error than model B.

    [*6 marks*]

    d) If a model has a zero training error and a non-zero test error, the model is overfitting.

    [*6 marks*]

    e) A model can be simultaneously underfitting and overfitting.

    [*6 marks*]

2. In neural networks, batch normalization is a commonly used operation where a set of variables are normalized before passed to subsequent computations. Formally, given a set (batch) of real values $x_1, \dots, x_B$, batch normalization returns a set of real values $y_1, \dots, y_B$ where

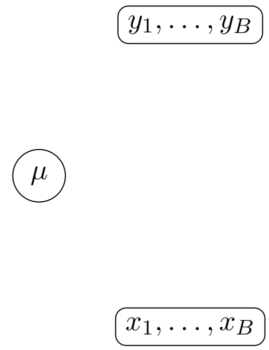
$$y_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

and

$$\mu = \frac{1}{B}\sum_{i=1}^{B} x_i \qquad \sigma = \sqrt{\frac{1}{B}\sum_{i=1}^{B} x_i^2 - \mu^2}. \tag{2}$$

If the loss function is L, we would like to compute the gradients through batch normalization. We are given $\frac{\partial L}{\partial y_i}$ for $i = 1, \dots, B$.

    a) Complete the following computation graph by drawing edges from input nodes to output nodes for each operation in batch normalization. There are a total 6 edges.



[6 marks]

b) Derive $\frac{\partial L}{\partial \sigma}$ based on the computation graph.

[8 marks]

c) Derive $\frac{\partial L}{\partial \mu}$ based on the computation graph. Note that σ depends on μ, and you do not need to substitute $\frac{\partial L}{\partial \sigma}$ with the answer in a).

[8 marks]

d) Derive $\frac{\partial L}{\partial x_j}$ for a particular $j \in \{1, \dots, B\}$. Note that y_j, μ, and σ depend on x_j, and you do not need to substitute $\frac{\partial L}{\partial \sigma}$ and $\frac{\partial L}{\partial \mu}$ with the answers in a) anb b).

[8 marks]

3. Gaussian mixture models (GMM) and k-means share a lot of similarities.

Given a data set $\{x_1, \dots, x_n\}$, GMM assumes that there is a hidden variable $z_i \in \{1, \dots, K\}$ for every data point x_i, where K is the number of Gaussian components. The mean for the k-th component GMM is μ_k and its variance is σ_k^2. The prior for choosing the k-th component is $v_k \in [0, 1]$ where $\sum_{i=1}^{K} v_i = 1$. Given the parameters, the distributions can be written as

$$p(x|z) = \frac{1}{(2\pi)^{d/2}|\Sigma_z|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_z)^\top \Sigma_z^{-1}(x - \mu_z)\right) \tag{3}$$

$$p(z) = v_z \tag{4}$$

The variational lower bound of the log likelihood is

$$L = \sum_{i=1}^{n} \left[\mathbb{E}_{z \sim q(z|x_i)}[\log p(x_i|z)] - \mathrm{KL}[q(z|x_i) \| p(z)]\right]. \tag{5}$$

The expectation-maximization optimizes L by iteratively updating GMMs with the update rules

$$q(z|x) \leftarrow p(z|x) \tag{6}$$

$$\mu_z \leftarrow \frac{\sum_{i=1}^{n} q(z|x_i) x_i}{\sum_{i=1}^{n} q(z|x_i)} \qquad \text{for } z = 1, \dots, K \tag{7}$$

$$\Sigma_z \leftarrow \frac{\sum_{i=1}^{n} q(z|x_i) x_i x_i^\top}{\sum_{i=1}^{n} q(z|x_i)} - \mu_z \mu_z^\top \qquad \text{for } z = 1, \dots, K \tag{8}$$

a) Show that L becomes $\sum_{i=1}^{n} \log p(x_i)$ if we let $q(z|x) = p(z|x)$.

[10 marks]

b) Show that L is concave in μ_z for $z = 1, \ldots, K$ when q is fixed. Note that when q is fixed, it no longer depends on μ_z.

[15 marks]

c) Use Bayes rule to derive $q(z|x)$ is terms of $p(x|z)$ and $p(z)$.

[5 marks]

For k-means, we have k mean vectors $\mu_1, \ldots, \mu_K$. The update rule for k-means is

$$z_i = \underset{k=1,\ldots,K}{\operatorname{argmin}} \|x_i - \mu_k\|^2 \qquad \text{for } i = 1, \ldots, n \tag{9}$$

$$\mu_k = \frac{\sum_{i=1}^{n} \mathbb{1}_{z_i=k} x_i}{\sum_{i=1}^{n} \mathbb{1}_{z_i=k}} \qquad \text{for } k = 1, \ldots, K \tag{10}$$

d) Ignoring the update of the variance, how would you change the GMM update rules so that they become k-means?

[10 marks]