Homework 3

Issue Date: March 20, 2019
Due Date: April 3, 2019

1. We explained in class that one of the most important assumptions in machine learning is that of the IID (independent and identically distributed) assumption of the training data. We also explained extensively how the independence assumption make the computations simple and tractable. Briefly explain what the *identically distribution* assumption mean and why it is important.

2. Given a choice between a vanilla linear regression ($f_\theta(x^{(i)}) = \theta_0 + \theta^T b(x^{(i)})$) and maximum likelihood regression with Gaussian mean noise ($p(y^{(i)}|x^{(i)};\theta) = \mathcal{N}(y^{(i)};\theta^T b(x^{(i)}),\sigma^2)$), which one will you choose? Explain the reasoning behind your choice.

3. The maximum likelihood estimate of the variance of a Gaussian distribution is given as $\sigma^2_{ML} = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{ML})^2$ where m is the number of examples. However, Octave/MathLab *var* function implements the variances as $\sigma^2_{ML} = \frac{1}{m-1}\sum_{i=1}^{m}(x^{(i)} - \mu_{ML})^2$. Explain whether the disparities between the two estimates have any effects on the accuracy of the model.

4. We spoke about the multivariate Gaussian distribution in class. Given a dataset $\mathcal{D} = \{x_1^{(1)}, x_2^{(2)}, ...x_n^{(m)}\}$ where each $x^{(i)}$ is an n-dimensional feature vector, give, in terms of n, the total number of parameters the multivariate Gaussian distribution fitted over the data will have if

   (a) we assume that the features $\{x_1^{(i)}, ..., x_n^{(i)}\}$ of each example are independent.

   (b) we ignore the assumption in (a) in modeling our distribution.

   **Note that, the IID assumption between the examples still hold in (a).

5. Despite the overwhelming merits of cross-validation over hold-out validation, hold-out validation is still the predominant method for hyper-parameters fine-tuning in deep learning. Can you explain briefly why this is the case?

6. Briefly explain, giving examples, the difference between hyper-parameters and parameters.

7. Explain briefly why the Gaussian distribution is a very good model for implementing anomaly detection? Can you think of instances where the Gaussian is not a good model to perform anomaly detection?

8. Explain the relationship between regularization parameter and model complexity.

9. Compare and contrast ML and MAP regression.

10. Why might you use Root Mean Square error instead of just error in depicting models?