UNIVERSITY OF WARWICK

CS342: Machine Learning

Sample questions for summer examinations 2019/20

---

1. Recall the $K-$means clustering algorithm, which assumes that there are K clusters and each $d-$dimensional feature vector, or point, is close to its cluster centre. Let $\mathbf{X}$ be a $n \times d$ matrix containing the n training points to be clustered, and $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \hat{y}_3, \ldots, \hat{y}_n]^T$ a one-column vector containing the values of the clusters assigned to the points. For example, $\hat{y}_i = 2$ indicates that point $\mathbf{x}_i$ is assigned to the second cluster with center $\mathbf{m}_2$. The objective function for $K-$means can be formulated in terms of cluster centers, $\{\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \ldots, \mathbf{m}_K\}$, and values of the clusters assigned, as follows:

$$f(\{\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \ldots, \mathbf{m}_K\}, \{\hat{y}_1, \hat{y}_2, \hat{y}_3, \ldots, \hat{y}_n\}) = \sum_{i=1}^{n} ||\mathbf{m}_{\hat{y}_i} - \mathbf{x}_i||^2 \tag{1}$$

where $\mathbf{m}_{\hat{y}_i}$ is the cluster center assigned to point $\mathbf{x}_i$.

   (a) Based on Eq. 1, define the objective function of a particular $\hat{y}_i$. [4]

   (b) Based on Eq. 1, define the objective function of a particular dimension j of the k^{th} center, $\mathbf{m}_k$; in other words, for a particular m_{kj}. [6]

   (c) Explain how the function in 1a can be minimized to optimize the cluster assignment. [4]

   (d) Using the function in 1b, mathematically show how to solve for m_{kj} to optimize the mean. [6]

   (e) If the objective function in Eq. 1 uses as distance the L1-norm instead, clearly explain what the two-step optimization process of the original $K-$means clustering algorithm optimizes in this case. [5]

---

2. A linear model for regression is given by $\hat{y}_i = w_1 x_i + w_0$ for one-dimensional feature vectors, where $\hat{y}_i$ is the predicted value for feature vector x_i, and (w_o, w_1) are the regression weights.

   (a) What is the expected error if the bias is equal to zero? Assume the model has a training error $= 0$ has been optimized based on quadratic errors. [5]

   (b) Clearly explain how the regression weights can be learned by using a change of basis if the training process assumes a model with no bias. [5]

   (c) Clearly explain, with an example, why least squares may not have a unique solution if two features are collinear. Assume two-dimensional feature vectors for your explanations. [5]

   (d) Clearly explain, with an example, how a linear model can be used for regression when the targets, $\{y_i\}$, are not a linear function of the feature vectors $\{\mathbf{x}_i\}$. In your explanations, assume two-dimensional feature vectors and a quadratic relationship between $\{y_i\}$ and $\{\mathbf{x}_i\}$. [5]

   (e) Within the context of polynomial regression with a change of basis, clearly explain the concept of overfitting and explain how overfitting can be reduced during optimization. Use a sample objective function in your explanations. [5]

3. (a) You are part of a team in charge of designing a machine-learning model to predict the occurrence of tornados based on d features. The model will be based on regression to first predict a numerical value, $\hat{y}_i$, for each d-dimensional feature vector, $\mathbf{x}_i$. The predicted numerical values, $\{\hat{y}_i\}$, will then be used to output the correct class, *tornado* or *no-tornado*. You are in charge of designing the regression model that will predict $\{\hat{y}_i\}$. You notice that most of the training feature vectors follow a linear trend with respect to the target values, $\{y_i\}$; however, those that correspond to the event of interest (i.e., a tornado), do not follow this linear trend. Moreover, there are a very limited number of training feature vectors that correspond to the event of interest, in other words, these feature vectors can be considered as outliers within the training data.

   i. Cleary explain the difference between using the squared of the L2-norm and the L1-norm to train the regression model. For each norm, give the objective function to use. Specify and justify which norm is more appropriate for this particular problem. [8]

   ii. Clearly explain if the infinity norm, L∞, can provide a more robust regression model for this particular problem. In your explanations, include the objective function to use for this norm. [8]

   (b) Consider a regression model for d-dimensional feature vectors whose L1-regularized objective function is of the form $f(\mathbf{w}) = \frac{1}{2}||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 + \lambda||\mathbf{w}||_1$, where $\mathbf{w}$ are the regression weights to be learned, λ is a hyperparameter that controls the sparsity of the solution, and $\mathbf{X}$ and $\mathbf{y}$ are the training samples and their target values, respectively. Minimizing the L1-regularized objective function, unfortunately, may not result in a unique solution. Clearly explain how to modify this objective function to guarantee a sparse and unique solution. In your explanations, show the modified objective function. [9]

4. Consider a learning model trained to predict numerical values, $\{\hat{y}_i\}$, for d-dimensional data points, $\mathbf{x}_i$.

   (a) Using diagrams similar to the ones used to depict the architecture of neural networks (NNs), graphically depict the learning model when it is designed as: [10]

      i. a simple linear regression model;
      ii. a linear regression model with manual change of basis (new basis has k-dimensional data);
      iii. a linear regression model with change of basis learned by a separate latent factor model (new basis has k-dimensional data);
      iv. a linear regression model with change of basis learned jointly with the regression weights (new basis has k-dimensional data); and
      v. a 1-layer NN with k neurons.

      For all diagrams, assume no bias is used and clearly define all variables used.

   (b) For each case listed in 4a, describe the parameters to be learned. Show the dimensions of the all parameters for $d = 3$ and $k = 3$. [10]

   (c) Clearly explain why the Sigmoid function is commonly used in NNs. In your explanations, use the equation of the function to support your answers. [5]