CS570 박사자격 시험

1. [20 points] Consider a linear model of the form

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

together with a sum-of-squares error function of the form

$$error_D(\mathbf{w}) = \frac{1}{2} \sum_{t=1}^{N} \{y(\mathbf{x}_t, \mathbf{w}) - r_t\}^2$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. Show that minimizing $error_D$ averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of some weight-decay regularization term.

2. [20 points] Consider a binary classification problem in which each observation $\mathbf{x}_n$ is known to belong to one of two classes, corresponding to $t = 0$ and $t = 1$, and suppose that the procedure for collecting training data is imperfect, so that training points are sometimes mislabeled. For every data point $\mathbf{x}_n$, instead of having a value t for the class label, we have instead a value π_n representing the probability that $t_n = 1$. Given a probabilistic model $p(t = 1|\phi(\mathbf{x}))$, write down the log likelihood function appropriate to such a data set.

3. [60 points] Consider Gaussian Mixture Model $p(\mathbf{x}|\boldsymbol{\theta}) = \sum_k \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, and the log likelihood $\ell(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\boldsymbol{\theta})$. Define the posterior responsibility that cluster k has for datapoint $\mathbf{x}_n$ as:

$$r_{nk} = p(z_n = k|\mathbf{x}_n, \boldsymbol{\theta}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}$$

   a. Derive formula for the gradient of the log likelihood w.r.t. $\boldsymbol{\mu}_k$: $d\ell(\boldsymbol{\theta})/d\boldsymbol{\mu}_k = \ldots$

   b. Do the same for $d\ell(\boldsymbol{\theta})/d\pi_k = \ldots$

   c. We need to fix the above formula since we need to impose the constraint $\sum_k \pi_k = 1$. This can be done by reparameterizing with the softmax function $\pi_k = \exp(w_k)/\sum_{k'} \exp(w_{k'})$. Now, derive the formula for the gradient $d\ell(\boldsymbol{\theta})/dw_k = \ldots$

   d. Derive formula for the gradient w.r.t. $\boldsymbol{\Sigma}_k$: $d\ell(\boldsymbol{\theta})/d\boldsymbol{\Sigma}_k = \ldots$

   e. We also need to fix the above formula since we need $\boldsymbol{\Sigma}_k$ to be symmetric positive definite. This can be done by reparameterizing with Choleskey decomposition $\boldsymbol{\Sigma}_k = \mathbf{R}_k^\top \mathbf{R}_k$ where $\mathbf{R}_k$ is an upper-triangular matrix. Derive the gradient w.r.t. $\mathbf{R}_k$: $d\ell(\boldsymbol{\theta})/d\mathbf{R}_k = \ldots$

   f. What are the advantages of EM over this gradient ascent approach?