Assignment (EECS6327 F19)

Due: in class on Dec 4, 2019.

You have to work individually. Hand in a hardcopy of your answers before the deadline. No late submission will be accepted. No handwritting is accepted. Direct your queries to Hui Jiang (hj@cse.yorku.ca).

1. (**Soft-margin Support Vector Machine**) The primary problem of soft-margin support vector machine (SVM) is

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\mathbf{w}^\intercal\mathbf{w} + C\sum_{i=1}^{N}\xi_i$$

subject to

$$y_i(\mathbf{w}^\intercal\mathbf{x}_i + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0 \quad \forall i \in \{1, 2, \cdots, N\}$$

use the Lagrangian technique to derive its dual problem as

$$\max_{\boldsymbol{\alpha}} \quad \mathbf{e}^\intercal\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\intercal\mathbf{Q}\boldsymbol{\alpha}$$

subject to

$$\mathbf{y}^\intercal\boldsymbol{\alpha} = 0$$

$$0 \leq \boldsymbol{\alpha} \leq C$$

2. (**Convolutional Neural Networks**) Consider a simple convolutional neural network consisting of two hidden layers, each of which is composed of convolution and ReLU, and then followed by a max pooling and soft-max layer. Assume each convolution uses K kernels of 5×5 with a stride of 1 in each direction (no zero-padding), each kernel is denoted as a tensor of $\mathbf{w}(f_1, f_2, p, k, l)$, where $1 \leq f_1, f_2 \leq 5$, $1 \leq k \leq K$, and l indicates the layer number $l \in \{1, 2\}$, and p indicates the number of features maps in each layer. The max pooling layer uses 4×4 patches with a stride of 4 in each direction.

(a) Derive the error back-propagation (BP) to compute the gradients for all kernels $\mathbf{w}(f_1, f_2, p, k, l)$ in this network when the cross-entropy loss is used.

(b) In object recognition, translating an image by a few pixels in some direction should not affect the category recognized. Suppose that we consider image with an object in the foreground on top of uniform background. Also suppose that the objects of interest are always at least 10 pixels away from the borders of the image. Is this neural network invariant to translations of at most 10 pixels in some direction? Here the translation is applied only to the foreground object while keeping the background fixed. If your answer is yes, show that the neural network will necessarily produce the same output for two images where the foreground object is translated by at most 10 pixels. If your answer is no, provide a counter example by describing a situation where the output of the neural network is different for two images where the foreground object is translated by at most 10 pixels. If your answer is no, can you find any particular translation by at most 10 pixels, where the neural network will generate invariant output only for this translation?

3. (**Transformer**) Suppose that we have a multi-head transformer, where $A^{(j)}, B^{(j)} \in \mathbb{R}^{d\times l}, C^{(j)} \in \mathbb{R}^{d\times o} \quad (j = 1 \cdots J)$, see lecture notes for the detailed structure of such a transformer.

(a) Estimate the computational complexity of the forward pass of this transformer for an input sequence $X \in \mathbb{R}^{n\times d}$.

(b) Derive the error back-propagation (BP) to compute the gradients for $A^{(j)}, B^{(j)}, C^{(j)}$ when an objective function $Q(\cdot)$ is used.

4. (**Maximum Likelihood Estimation**) Assume we have K different classes, i.e. $\omega_1, \omega_2, \cdots, \omega_K$. Each class $\omega_k \quad (k = 1, 2, \cdots, K)$ is modeled by a multivariate Gaussian distribution with the mean vector $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}$, i.e., $p(\mathbf{x} \mid \omega_k) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the common covariance matrix for all K classes. Suppose we have collected N data samples from these K classes, i.e., $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, and let $\{l_1, l_2, \cdots, l_N\}$ be their labels so that $l_n = k$ means the data sample $\mathbf{x}_n$ comes from the k-th class, ω_k.

Based on the given data set, derive the maximum-likelihood estimates for all model parameters, i.e., all mean vectors $\boldsymbol{\mu}_k$ $(k = 1, 2, \cdots, K)$ and the common covariance matrix $\boldsymbol{\Sigma}$.

5. (**EM algorithm**) Consider a D-dimensional variable $\mathbf{x}$, each of whose dimensions, x_d, is an integer. Suppose the distribution of these variables is described by a mixture of the multinomial distributions so that

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}|\boldsymbol{\mu}_k) \propto \sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} \mu_{kd}^{x_d}$$

where the parameter μ_{kd} denotes the probability of d-th dimension in k-th component, subject to $0 \leq \mu_{kd} \leq 1$ $(\forall k, d)$ and $\sum_d \mu_{kd} = 1$ $(\forall k)$.

Given an observed data set $\{\mathbf{x}_n\}$, where $n = 1, \cdots, N$, derive the E and M step equations of the EM algorithm for optimizing the mixing weights π_k $(\sum_k \pi_k = 1)$ and the component parameters μ_{kd} of this distribution by maximum likelihood.