Examen Apprentissage Statistique - 7/01/2021
Partie 2 : Réseaux de Neurones – Deep Learning
Master Probabilités Finance – Sorbonne Université -Ecole Polytechnique
Notes de cours autorisées- durée 1h30

I. Course

Provide a synthetic description of GANs: objective, algorithm principles and some examples of applications (max 1 page).

II. QCM GANs

Answer Y/N just before the question number please.

1. A GAN allows us to learn any distribution.
2. The main applications of GANs are for text generation.
3. Inference in GANs amounts at sampling from the latent space ($variable\ z$) and then computing the output of the generator in a deterministic way.
4. The training algorithm for GAN is a stochastic gradient descent algorithm used for optimizing an approximation of the data likelihood.
5. The GAN discriminator optimizes a cross-entropy criterion.
6. If for a given generator, the discriminator perfectly separates the ground truth data and the simulated data, the gradient w.r.t. the parameters of the discriminator is 0.
7. Let us suppose that the GAN has reached its equilibrium, then the optimal discriminator computes $D(x) = \frac{1}{2}, \forall x$.
8. The distribution of the latent variables (z) is usually chosen as a simple distribution (Gaussian, uniform, etc).
9. The objective of GAN is to generate a distribution that is indistinguishable from the distribution of the observed data..
10. In order to train the GANs one samples from the space of latent variables (z) and from the space of data (x).

III. Exercise

We consider a Bayesian approach for supervised learning problems (classification or regression), implemented with a neural network (NN). We will examine how introducing simple constraints will allow us to control the complexity of a trained model. Let us denote $\boldsymbol{w}$ a random variable associated to the weight vector of a NN.

1. Our objective is to maximize the posterior (MAP) : $p(\boldsymbol{w}|X,Y) \propto p(Y|X,\boldsymbol{w})p(\boldsymbol{w})$

where $D = (X,Y)$ corresponds to the training set. One considers the following loss:

$$L_R(\boldsymbol{w}) = -\ln p(Y|X,\boldsymbol{w}) - \ln p(\boldsymbol{w}) = L(\boldsymbol{w}) + R(\boldsymbol{w})$$

Note that $L_R(\boldsymbol{w}) \propto -\ln p(\boldsymbol{w}|X,Y)$

Let us suppose that the prior distribution on the weights follows a Gaussian $p(\boldsymbol{w};\boldsymbol{0},\sigma) = \mathcal{N}(\boldsymbol{w};\boldsymbol{0},\sigma^2 I)$ with $\boldsymbol{0}$ the mean vector, σ a positive scalar and I the identity matrix.

1.1. Show that $R(\boldsymbol{w})$ writes as $C + \lambda \boldsymbol{w}^T\boldsymbol{w}$, $\lambda \in R^+$, C a constant.
1.2. What is the effect of this term on the weight vector $\boldsymbol{w}$, how does it allows us controlling the model complexity?

2. One considers now a NN with one hidden cell layer, the activations of the hidden layer are non-linear, and the output layer activations are linear. One would like the regularization penalty to

provide consistency properties for elementary transformations of the data. Here consistency means that the computed output of the NN does not change under these transformations and that the result of the training does not change either. For example, let us consider a linear transformation of the inputs $x' = ax$, a simple transformation $w'^1 = \frac{1}{a}w^{(1)}$ with $w^{(1)}$ the weight vector of the first layer will provide an unchanged result. Similarly, the output transformation $\hat{y}' = by$ will be compensated with $w'^{(2)} = bw^{(2)}$ with $w^{(2)}$ the weight vector of the second weight layer.

2.1. Is the penalization introduced in question 1 consistent w.r.t. these two transformations $w'^1 = \frac{1}{a}w^{(1)}$ and $w'^{(2)} = bw^{(2)}$, in the sense that training will lead to the same result? One could for example examine if the penalization term $\lambda w^T w$ remains unchanged under these weight transformations.

2.2. Let us consider the prior $p(w) = p(w^{(1)}, w^{(2)}) = p(w^{(1)})p(w^{(2)})$ with $p(w^{(1)}) = \mathcal{N}(w^{(1)}; 0, \sigma_1 I)$ and $p(w^{(2)}) = \mathcal{N}(w^{(2)}; 0, \sigma_2 I)$. Give the expression of the corresponding regularization term $R(w) = -\ln p(w)$ and show that it can write as $\lambda_1 w^{(1)T} w^{(1)} + \lambda_2 w^{(2)T} w^{(2)}$

2.3. Is this prior consistent w.r.t the above transformations? As before, one could evaluate if this prior changes with the corresponding weight transformations.

3. Let us now introduce a prior distribution on the weights under the form of a Gaussian mixture. The distribution of variable w is $p(w) = \prod_i p(w_i)$, with $p(w_i) = \sum_{j=1}^m \pi_j \mathcal{N}(w_i; \mu_j, \sigma_j^2)$, w_i is a scalar component of w, m is the number of components in the mixture, $\mathcal{N}(w_i; \mu_j, \sigma_j^2)$ is a **one dimensional** Gaussian distribution, with mean μ_j and variance σ_j^2, the π_j are the mixture coefficients, they verify $\sum_{j=1}^m \pi_j = 1$. One now considers the following regularization term :

$$R(w) = -\sum_i \ln\left(\sum_{j=1}^m \pi_j \mathcal{N}(w_i; \mu_j, \sigma_j^2)\right)$$

and the loss function : $L_R(w) = L(w) + \lambda R(w)$

where $L(w)$ is a classical loss defined on the training set (cross entropy or mean square error for example) and $\lambda \in R^+$.

3.1. Draw a sketch of a mixture model with three components, with mixture coefficients chosen as you wish.

3.2. Let us suppose that all the parameters $\mu_j, \sigma_j^2, \pi_j, j = 1 \dots m$ are fixed, one minimizes $L_R(w)$ by optimizing the weights. What is the effect of the constraint on the weights?

3.3. Training concerns all the model parameters: the weights w and the mixture parameters μ_j, $\sigma_j^2, \pi_j, j = 1 \dots m$. Let us consider the following posterior probability, $p(j|w) = \gamma_j(w) = \frac{\pi_j \mathcal{N}(w;\mu_j,\sigma_j^2)}{\sum_{k=1}^m \pi_k \mathcal{N}(w;\mu_k,\sigma_k^2)}$. Beware, here w is a scalar variable.

3.3.1. Write the expression of $\frac{\partial R}{\partial w_l}$ as a function of the $\gamma_j(w)$s. What is the effect of the regularization term on the weights?

3.3.2. Write the expression of $\frac{\partial R}{\partial \mu_j}$ as a function of the $\gamma_j(w)$s. What is the effect of the regularization term on the μ_js?

3.3.3. Write the expression of $\frac{\partial R}{\partial \sigma_j}$ as a function of the $\gamma_j(w)$s. What is the effect of the regularization term on the σ_js? The coefficients σ_j shall remain positives, propose a solution for that.