Final Exam: Deep Learning: Spring 2014

Yann LeCun

May 19, 2014

1. (35 points) **General questions:**
   (a) (5 points) What is pooling, where is it used, and what is its purpose?
   (b) (5 points) what is a recurrent net?
   (c) (5 points) how do we apply backprop to a recurrent net?
   (d) (5 points) what is metric learning?
   (e) (5 points) give the formula for (and the name of) a common metric learning criterion.
   (f) (5 points) Structured prediction is used whenever the output has "structure", i.e. the output is composed of multuple items that must respect certain rules or constraints. Give examples of applications where structured prediction would be used.
   (g) (5 points) You are given a dataset with 1 million samples, each sample has 10,000 variables, some of which are counts (e.g. how many times people have clicked on a particular link, visited a particular page, or bought a particular product), others are measurements whose distributions seem close to Gaussian. Describe the steps to prepare the data.
   (h) (5 points) You are given X-ray images of patients where malignant tumors have been labeled at the pixel level (for each X-ray, you have another image where each pixel is labeled 1 for tumor and -1 for non tumor). Describe how you could use deep learning to solve the problem of detecting malignant tumors. specify which type of model you would use, and how you would go about setting up the dataset to train it.

2. (10 points) `sort` **module:** The 2-2 `sort` module takes two values x_1 and x_2 and outputs two values $z_1 = \max(x_1, X_2)$ and $z_2 = min(x_1, x_2)$
   (a) (5 points) assuming we know $z'_1 = \frac{\partial L}{\partial z_1}$ and $z'_2 = \frac{\partial L}{\partial z_2}$, what are $x'_1 = \frac{\partial L}{\partial x_1}$, $x'_2 = \frac{\partial L}{\partial x_2}$.
   (b) (5 points) Given an N-dimensional input vector X, the linear module performs the matrix-vector multiplication $Z = WX$. The rotation module rotates the input $Z = R(X)$. The permutation module permutes the components of the input $Z_i = X_{\rho(i)}$ where ρ is a permutation over the integers $[1, N]$. The sort module outputs the components of X sorted in the descending order. Which of these modules are special cases of the others (i.e which ones could in principle be implemented with the same code, if it weren't for efficiency issues)?

3. (30 points) **Softmax:**
   The N-N `softmax` module takes an N-dimensional vector Z and outputs another N-dimensional vector Q:

$$Q_i = \frac{\exp(-\beta Z_i)}{\sum_{k=1}^{N} \exp(-\beta Z_k)}$$

(a) (10 points) compute the backprop formula, i.e. give the expression of $\frac{\partial L}{\partial Z_u}$ as a function of the quantities $\frac{\partial L}{\partial Q_v}$. Note: there are two cases: $u = v$ and $u \neq v$.

(b) (5 points) Consider the following two learning machines designed to perform classification. Machine A is a paramterized function $Z = G(X, W)$ (e.g. a neural net) trained with the negative log-likelihood loss:

$$L_{NLL} = Z_k + \frac{1}{\beta} \log \left[\sum_{k=1}^{N} \exp(-\beta Z_k) \right]$$

Machine B is the same function $Z = G(X, W)$ followed by a `softmax` module. Is there a loss function applied to machine B that makes it identical to machine A. Which one?

(c) (5 points) The Kullback-Leibler divergence between two probability distributions P and Q can be used as a loss function in which Q is the probability distribution produced by a machine, and P is the desired distribution $L_{KL} = \sum_{i=1}^{N} P_i \log \frac{P_i}{Q_i}$. Show that there is a special set of distributions Ps for which L_{KL} is equivalent to L_{NLL}.

(d) (10 points) give the backprop formula, i.e. $\frac{\partial L}{\partial Z_u}$ as a function of P and Q (or Z).

4. (5 points) **Shared Weights:** Consider N neural nets $G(X_k, W_k) \quad k \in [1, N]$, where the X_k are a bunch of imput vectors, and the W_k a bunch of parameter vectors. The outputs of the networks are combined and fed to a loss function L.

(a) (5 points) As it turns, we want all the networks to share the same weight vector $W_k = W \ \ \forall k$. Assuming that one can compute the individual $\frac{\partial L}{\partial W_k}$, what is $\frac{\partial L}{\partial W}$?

5. (10 points) **ConvNet:** A 1D convolution layer has the form $Z_i = \sum_{j=0}^{K-1} W_j X_{i+j}$ for $i \in [1, N]$, where K is the kernel size.

(a) (10 points) Assuming we know $\frac{\partial L}{\partial Z_i}$ for all i, what is the formula for $\frac{\partial L}{\partial W_j}$. Show that it actually is the result of a convolution.

6. (30 points) **Energy-Based Learning:** In energy-based learning, the system uses a energy function $F(X, Y, W)$ where X is the observed input, Y is the output to be predicted, and W is the trainable parameter.

(a) (5 points) Inference consists in finding "good" values for Y. How is this done?

(b) (5 points) Learning consists in finding a W that gives the right "shape" to the energy function F. This is done by minimizing a loss function $L(W, X^i, Y^i)$ averaged over training samples $(X^i, Y^i) \quad i \in [1, P]$. Write (in English) the property that a loss function should have, in terms of the energies of the various possible outputs.

(c) (10 points) Assuming that Y is a discrete variable, give three possible loss functions expressed as functions of the energies of the possible outputs and the desired output.

(d) (10 points) If the system has latent variables, the energy is denoted $E(X, Y, Z, W)$ where Z is the set of latent variables. Give two possible expressions for $F(X, Y, W)$ as a function of $E(X, Y, Z, W)$.

7. (20 points) **Sparse Coding:** Given an $N \times K$ dictionary matrix W_d, sparse coding represents a particular N-dimensional input vector Y as the K-dimensional code vector Z^* that minimizes

the energy function

$$E(Y, Z, W_d) = ||Y - W_d Z||^2 + \alpha \sum_{k=1}^{K} |Z_k|$$

$$Z^* = \text{argmin}_Z E(X, Z, W_d)$$

(a) (5 points) Generally, K is set larger than N. Why?

(b) (5 points) The ISTA and FISTA algorithms have the form $Z(t+1) = G(W_e.Y + SZ(t))$. What is the function G?

(c) (10 points) Give the expressions of W_e and S as a function of W_d, and α.

8. (10 points) **Auto-Encoders:** auto-encoders are machines of the form

$$\tilde{Y} = \text{Decoder}(W_d, Z) \quad Z = \text{Encoder}(W_e, Y))$$

where W_e are the parameters of the encoder, and W_d the parameters of the decoder.

(a) (5 points) Given a training set $\{Y^i, i \in [1, P]\}$, write a possible loss function with which to train a *sparse* auto-encoder.

(b) (5 points) Assuming the decoder is linear $\text{Decoder}(W_d, Z) = W_d Z$, what constraint should we add to prevent the system from converging to a trivial and useless solution?

9. (20 points) **Optimization:** A 2-input linear regressor is trained with square loss (Mean Squared Error):

$$\mathcal{L}(W) = \frac{1}{P} \sum_{i=1}^{P} \frac{1}{2} ||Y^i - W^T X^i||^2$$

Variable 1 has a mean 3 and variance 4, while variable 2 has mean -4 and variance 1 and is otherwise uncorrelated with variable 1.

(a) (10 points) Write the Hessian of $\mathcal{L}(W)$.

(b) (10 points) what can you do the data to make the Hessian equal to the identity matrix?

10. (10 points) **Optimization in multi-layer nets:** Consider the following linear auto-encoder with 1 input and 1 output: $\tilde{x} = w_2 w_1 x$, trained with the squared reconstruction error:

$$\mathcal{L}(W) = \frac{1}{P} \sum_{i=1}^{P} \frac{1}{2} (x^i - w_2 w_1 x^i)^2$$

The scalar training samples have variance 1.

(a) (5 points) What is the set of solutions (with 0 loss)?

(b) (5 points) Does the loss have a saddle point? Where?