Trial Midterm

Special Topics in Advanced Machine Learning
Spring 2017
Instructor: Anna Choromanska

Problem 1 (50 points)

The Kullback-Leibler (KL) Divergence measures how different two distributions $P(x)$ and $Q(x)$ are and is given as:

$$D(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

Evaluate the KL-divergence between the two Poisson distributions below:

$$P(x) = \frac{(\lambda_1)^x e^{-\lambda_1}}{x!}$$

$$Q(x) = \frac{(\lambda_2)^x e^{-\lambda_2}}{x!},$$

where x takes values $0, 1, 2, \ldots$ and λ_1 and λ_2 denote the means of respectively the first and the second distribution.

Problem 2 (70 points)

Suppose we have a box containing 8 apples and 4 oranges and a second box containing 10 apples and 2 oranges. One of the boxes is chosen at random (with equal probability of choosing either box) and an item is selected from the box and found to be an apple. Use Bayes' rule to find the probability that the apple came from the first box.

Problem 3 (130 points)

Assume we are given N pairs of data points $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, where each x and y is just a scalar and we wish to do a simple 1-dimensional linear regression with the function below:

$$f(x) = \theta_0 + \theta_1 x.$$

Assume we have the means of both the x and y, i.e.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i; \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

We find the optimal setting of θ_0^* and θ_1^* by minimizing the squared error:

$$L(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \theta_0 - \theta_1 x_i)^2.$$

a) (40 points) Put an "X" besides each statement that is true when we have the optimal least squared error setting for our parameters θ_0^* and θ_1^*:

( ) $\frac{1}{N}\sum_{i=1}^{N}(y_i - \theta_0^* - \theta_1^* x_i)y_i = 0$

( ) $\frac{1}{N}\sum_{i=1}^{N}(y_i - \theta_0^* - \theta_1^* x_i)(y_i - \bar{y}) = 0$

( ) $\frac{1}{N}\sum_{i=1}^{N}(y_i - \theta_0^* - \theta_1^* x_i)(x_i - \bar{x}) = 0$

( ) $\frac{1}{N}\sum_{i=1}^{N}(y_i - \theta_0^* - \theta_1^* x_i)(\theta_0^* + \theta_1^* x_i) = 0$

b) (90 points) Suppose we have the following components of the Gaussian sufficient statistics from the data. Show how we could compute the optimal value of θ_1^* only by using two of the following 5 scalar numbers:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i; \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$C_{xx} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2; \; C_{yy} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2; \; C_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}).$$

Problem 4 (100 points)

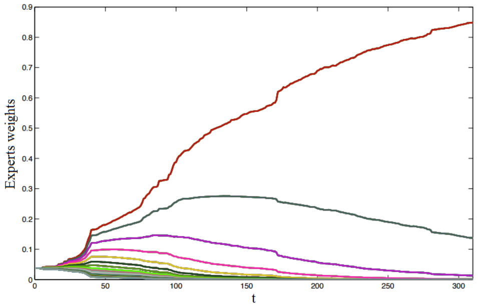
Consider the fragment of the convolutional architecture given below:

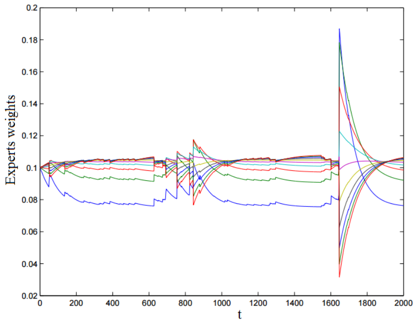
- Input image: $1 \times x \times y$
- Convolutional layer: $\underbrace{1 \to 4}_{\text{number of input and output channels}}, \underbrace{3 \times 4}_{\text{filter size}}, \underbrace{2 \times 2}_{\text{stride}}$
- ReLU
- MaxPooling: $\underbrace{2 \times 2}_{\text{region size}}, \underbrace{2 \times 2}_{\text{stride}}$
- Convolutional layer: $4 \to 6, 3 \times 3, 2 \times 2$
- ReLU
- MaxPooling: $2 \times 2, 2 \times 2$
- Flattening (3D to 1D):
$$\underbrace{6 \times 9 \times 6}_{\text{number of feature maps } \times \text{ size of the feature map } (9 \times 6)} \to 324$$

What is the size of the input (in other words what is x and y)?

Problem 5 (50 points)

Consider prediction with experts advice and two plots below showing the evolution of weights over different experts (that are color-coded) over time. Which of these two plots corresponds to the static-expert setting and which one corresponds to the fixed-share (α) setting and why?





Explain how you understand static-expert and fixed-share (α) settings. Write as much as you know about both settings.

Problem 6 (150 points)

Explain generative adversarial networks (GAN): i) describe the framework, ii) discuss the loss function and iii) discuss the algorithm for training GAN.

Problem 7 (120 points)

Put the phrases below in the correct places in the table:

a) identity of the best expert can change with time

b) the distribution over the experts is updates as $p_t(i) = \frac{1}{Z_t} p_{t-1}(i) e^{-\eta L_{t-1}(i)}$, where Z_t is a normalization factor, i is the i^{th} expert, η is the learning rate, and $L_{t-1}(i)$ is the value of the loss incurred by expert i at time t

c) stationary sequence

d) non-stationary sequence

e) identity of the best expert cannot change with time

f) experts share a fixed fraction of their weights with each other - this guarantees that the ratio of the weight of any expert to the total weight of all the experts may be bounded from below

Static-expert setting	Fixed-share (α) setting

Problem 8 (80 points)

Put the phrases below in the correct places in the table:

a) Endless stream of data

b) Memory available is $o(N)$ (e.g. $\sqrt{N}$)

c) Tested only at the very end

d) More than one pass may be possible

e) Fixed amount of memory

f) Each point is seen only once

g) Tested at every time step

h) Stream of (known and typically big) length N

Online model of computations	Streaming model of computations