Stat 232A-CS266A Homework #2 (total 7.5 points)

Due: Feb 19 Friday, upload to CCLE.

Problem 1 (Minimax entropy learning, 3 points). This question aims to refresh the proof process in minimax entropy learning. Let $p(\mathbf{I})$ be a FRAME model with K histograms matched to the underlying frequency $f(\mathbf{I})$

$$p(\mathbf{I};\Theta) = \frac{1}{Z(\Theta)} \exp\{-\sum_{i=1}^{K} < \lambda_i, H_i(\mathbf{I}) >\} \tag{1}$$

The parameter $\Theta = (\lambda_1, ..., \lambda_K)$ is learned so that the following constraints are satisfied.

$$E_p[H_i(\mathbf{I})] = E_f[H_i(\mathbf{I})] = h_i, \quad i = 1, 2, ..., K. \tag{2}$$

- Derive that

$$\frac{\partial \log Z}{\partial \lambda_i} = -E_p[H_i(\mathbf{I})].$$

- Let $\ell(\Theta)$ be the log-likelihood for one observed image $\mathbf{I}^{\text{obs}}$, prove that

$$\begin{aligned} \frac{\partial^2 \ell(\Theta)}{\partial \lambda_i \partial \lambda_j} &= -\frac{\partial^2 \log Z}{\partial \lambda_i \partial \lambda_j} && (3) \\ &= -E_p[(H_i(\mathbf{I}) - h_i)(H_j(\mathbf{I}) - h_j)], i, j \in \{1, 2, ..., K\} && (4) \end{aligned}$$

comment: Thus the second derivative of $\ell(\Theta)$ is a negative covariance matrix. So $\ell(\Theta)$ has a single maximum solution.

Now suppose we add a new feature from a dictionary $F_+ \in \Delta$, and augment the model to

$$p_+(\mathbf{I};\Theta_+) = \frac{1}{Z(\lambda_+)} \exp - \sum_{\alpha=1}^{K} < \lambda^*_\alpha, H_\alpha(\mathbf{I}) > - < \lambda_+, H_+(\mathbf{I}) > \tag{5}$$

The new parameter $\Theta_+ = (\lambda^*_1, ..., \lambda^*_K, \lambda_+)$ is learned so that p_+ satisfies the K constraints in eqn (2) plus one extra condition,

$$E_{p_+}[H_+(\mathbf{I})] = E_f[H_+(\mathbf{I})] = h_+. \tag{6}$$

Note that the previous parameters must be updated ($\lambda_\alpha \to \lambda^*_\alpha, i = 1, 2, ..., K$) when we introduce new features (marginal) because all features are correlated.

- Derive the steps for proving the following theorem

$$KL(f||p) - KL(f||p_+) = KL(p_+||p).$$

Question 2 (Learning by information projection, 2.5 points) Suppose that we are learning an underlying probability $f(\mathbf{I})$ for image $\mathbf{I}$, we start with an initial probability model $q(\mathbf{I})$, and observe that $q(\mathbf{I})$ has a different marginal probability over a macroscopic feature $H_i(\mathbf{I})$,

$$E_q[H_i(\mathbf{I})] \neq E_f[H_i(\mathbf{I})] = h_i.$$

where h_i is estimated from a set of examples sampled from $f(\mathbf{I})$. To improve the model, we learn a new probability model $p(\mathbf{I})$ so that it reproduces this margianl statistics. We denote the set of models that satisfy this constraint equation by,

$$\Omega_p = \{p : E_p[H_i(\mathbf{I})] = E_f[H_i(\mathbf{I})] = h_i.\}$$

Now, among all the $p(\mathbf{I})$ in Ω_p, we choose one that is closest to $q(\mathbf{I})$ so that it preserves the learning history.

$$p^* = \arg\min_{p \in \Omega_p} KL(p||q) = \arg\min_{p \in \Omega_p} \int p(\mathbf{I}) \log \frac{p(\mathbf{I})}{q(\mathbf{I})} d\mathbf{I}.$$

1. Derive the formula of $p(\mathbf{I})$ by the Euler-Lagrange equation (i.e. constrained optimization as we did for maximum entropy).
2. Prove that $KL(f||q) - KL(f||p) = KL(p||q)$. (Remark: Since $D(p||q) > 0$, p is closer to f than q).
3. Show that this is the maximum entropy principle when $q(\mathbf{I})$ is a uniform distribution.

Question 3 (Typical set, 2 points) Suppose we toss a coin N times and observe a 0/1 sequence (for head and tail respectively),

$$S_N = (x_1, x_2, ..., x_N), \quad x_i \in \{0, 1\}.$$

S_N is said to be of *type* q (i.e. the frequency of 1 is q in the sequence) with $q = \frac{1}{N} \sum_{i=1}^{N} x_i$.

Let $\Omega(q)$ be the set of all sequences S_N of type q. For simplicity, we discretize q to finite precision.

1. What is the cardinality of $\Omega(q)$ for $q = 0.2$ and $q = 0.5$ respectively? (Suppose we only care about the exponential order or rate).
2. Suppose we know that the underlying probability is $x_i = 1$ (or $x_i = 0$) with probability p (or $1 - p$ respectively), by sampling from this probability N times, what is the probability $p(S_N)$ that we observe a sequence $S_N \in \Omega(q)$? What is the total probability mass $p(\Omega(q))$ for all the sequences in set $\Omega(q)$?
3. In the above question, show that as $N \to \infty$,only sequences from the type p, i.e. set $\Omega(p)$, can be observed.