Problem 1 – Maximum Entropy Discrimination (MED)

Reading: Section 6 of Lecture 8.

In this problem, you will derive a MED classifier. Notation: to be consistent with the course notes, \( x_i \) will denote example \( i \), and \( x^j_i \) will denote the \( j \)th coordinate of example \( i \), and \( a^j \) will be the \( j \)-th coordinate of a vector \( a \in \mathbb{R}^n \). The number of samples is \( N \).

We consider the family of classifiers

\[
F = \{ f_w(x) = w^T x \mid w_{1:n} \in [-W, W] \}
\]  

(1)

The data domain is \( \mathbb{R}^n \), with \( x = [x_1 x_{n-1} 1] \). In other words, we add a constant coordinate to make \( f_w \) implement an affine classifier.

In MED, we search for a distribution \( q \) over \( F \) that has maximum entropy subject to classifying the data correctly. Since \( F \) is parametrized by \( w \in [-W, W]^n \), our \( q \) will be a distribution over \( [-W, W]^n \).

1. Write the MED primal for the domain we are considering. Denote the dual variables/Lagrange multipliers with \( \lambda_i \), \( i = 1 : N \). Write the expression of \( L(q, \lambda) \).

2. Compute the partial derivative \( \frac{\partial L}{\partial q(w)} \) and from it show that

\[
q(w) \propto e^{\sum_i \lambda_i y_i w^T x_i}
\]  

(2)

3. Denote \( A = \sum_i \lambda_i y_i x_i \). Show that the normalization constant for \( q(w) \) in (2) is

\[
Z(\lambda) = \prod_{j=1}^n \frac{e^{A^j W} - e^{-A^j W}}{A^j}
\]

(3)

4. Calculate the expression of \( \frac{\partial \ln Z(\lambda)}{\partial \lambda_i} \). Show that it is of the form \( y_i x_i^T B(\lambda) \) with \( B \in \mathbb{R}^n \).

5. Calculate the expression of \( \frac{\partial^2 \ln Z(\lambda)}{\partial \lambda_i \partial \lambda_k} \). Show that it is of the form \( y_i y_k x_i^T C(\lambda)x_k \) with \( C \) a diagonal \( 3 \times 3 \) matrix.

6. Write the dual MED problem as a convex minimization problem.
7. Compute the expected $w, \bar{w} = E_{q(\lambda)}[w]$, as a function of $\lambda$, (it will be a function of quantities you already have).

8. Write the expression of the MED classifier (it will be a function of quantities you already have), and show that it is a linear classifier.

9. Show that the expression of the negative entropy is

$$-H(q) = A^T \bar{w} - \ln Z(\lambda)$$

(There is more than one way to obtain this expression)

**Problem 2 – The Markov chain as maximum entropy model**

Let $X_1, X_2, \ldots X_n \in \{0, 1\}$ be a collection of binary random variables. We want to define a joint distribution $P$ over $X_{1:n}$ which has prescribed marginals, and maximum entropy.

2.1 Denote $x = (x_1:n) \in \{0, 1\}^n$ a configuration of $X$, by $p_x = P(X_{1:n} = x)$ the value of the joint distribution at $x$, and by $p$ the vector $p = (p_x)_{x \in \{0, 1\}^n}$. The above problem can be formulated as

$$\max_p H(p) \quad \text{s.t.} \quad E[X_i X_{i+1}] = r_i, \; i = 1 : n - 1 \quad (6)$$
$$E[X_i] = r_0 \quad (7)$$
$$E[1] = 1 \quad (8)$$

Write the problem explicitly as a function of $p$ and show that this problem is/can be formulated as a convex program.

2.2 Denote by $\theta_i, \theta_0, \theta'$ the Lagrange multipliers associated respectively with (6), (7), (8). Write the Lagrangean function $L(p, \theta_i, \theta_0, \theta')$ for this optimization problem, then calculate the value $p(\theta)$ that achieves $\inf_p L(p, \theta)$, for $\theta = (\theta_i, \theta_0, \theta')$.

Feel free to leave any normalization constant unevaluated (e.g replace it with $Z(\theta)$).

2.3 Examine the form of the maximum entropy solution $p(\theta)$ obtained in 2.2. Show that it represents a Markov chain.

[Problem 3 – $\nu$-SVN Not graded]

3.a Find the dual for the $\nu$-SVM Support Vector Machine (see also lecture notes)
defined by

\[
\begin{align*}
\text{minimize}_{w,b,\xi,\rho} & \quad \frac{1}{2}||w||^2 - \nu \rho + \frac{1}{N} \sum_i \xi_i \\
\text{s.t.} & \quad y_i(w^T x^i + b) \geq \rho - \xi_i \\
& \quad \xi_i \geq 0 \\
& \quad \rho \geq 0
\end{align*}
\tag{9}
\]

where \( \nu \in [0,1] \) is a parameter. 3.b Prove the following statements (also from the lecture notes) If \( \rho > 0 \) then:

- \( \nu \) is an upper bound on \( \# \text{margin errors}/N \) (if \( \sum_i \alpha_i = \nu \))
- \( \nu \) is a lower bound on \( \#(\text{original support vectors + margin errors})/N \)
- \( \nu \)-SVM leads to the same \( w, b \) as C-SVM with \( C = 1/\nu \)