

Statistical Methods for Artificial Intelligence
(TTIC 103, CMSC 35420)
Autumn 2007
Midterm Exam

1. Use Jensen's inequality to prove $KL(P, Q) \geq 0$ where $KL(P, Q) = E_{x \sim P} \left[\ln \frac{P(x)}{Q(x)} \right]$.

2. Consider the probability density $p(x)$ with $x \in \mathbb{R}^2$ defined as follows.

$$p(x) = \begin{cases} \frac{1}{\pi} & \text{if } x_1^2 + x_2^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

a. What is the mean vector and covariance matrix of this density?

b. Are x_1 and x_2 independent in the sense $p(x) = p_1(x_1)p_2(x_2)$ where p_1 and p_2 are the marginals of p ? Justify your answer.

3. Consider the "quadratic hinge loss" defined as follows.

$$L_{\text{hinge2}}(m) = \begin{cases} (m - 1)^2 & \text{for } m \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Define w^* as follows.

$$w^* = \underset{w}{\operatorname{argmin}} \sum_{t=1}^N L_{\text{hinge2}}(y_t(w \cdot \Phi(x_t))) + \frac{1}{2} \lambda \|w\|^2 \quad (1)$$

a. Is the optimization problem defined by (1) convex? Justify your answer.

b. Rewrite (1) as an optimization problem on α where we define $w = \sum_{t=1}^N \alpha_t \Phi(x_t)$. The optimization problem must be defined entirely in terms of the kernel function $K(x, y) = \Phi(x) \cdot \Phi(y)$ — you must assume that $\Phi(x)$ cannot be computed but that $K(x, y)$ can be computed.

c. Is the optimization problem you defined in part b. convex in α ? Justify your answer.

4. Suppose we want to model a probability density on \mathbb{R}^2 . We are given a data set x^1, \dots, x^N with $x^t \in \mathbb{R}^2$, or equivalently, $x^t = \langle x_1^t, x_2^t \rangle$. We define a model as follows where π is a $K \times K$ matrix, and μ and σ are K -dimensional arrays and the latent information z is a pair $\langle k, m \rangle$ with $k \in \{1, \dots, K\}$ and $m \in \{1, \dots, K\}$.

$$\Theta = \langle \pi, \mu, \sigma \rangle$$

$$P_{\Theta}(x^1, \dots, x^N, z^1, \dots, z^N) = \prod_{t=1}^N \pi_{z_1^t, z_2^t} \mathcal{N}(\mu_{z_1^t}, \sigma_{z_1^t})(x_1^t) \mathcal{N}(\mu_{z_2^t}, \sigma_{z_2^t})(x_2^t)$$

In this model the marginal on x_1 is a mixture of one dimensional Gaussian clusters and the marginal on x_2 is a different mixture of the same set of Gaussians. But the choice of the cluster for x_1 can be correlated with the choice of cluster for x_2 . The model stores a matrix π of the joint probability of the cluster selection.

We want to approximately solve the following optimization problem.

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} P_{\Theta}(x^1, \dots, x^N)$$

Assume we are given Θ_{OLD} . Let Θ_{NEW} be the result of a single EM update from Θ_{OLD} . Give a way of computing Θ_{NEW} from the data set and Θ_{OLD} . Your answer can be given in terms of the following quantities.

$$P_{k,m}^t = P_{\Theta_{OLD}}(z^t = \langle k, m \rangle \mid x^t)$$

You do not need to give the formula for computing $P_{k,m}^t$, just an expression for Θ_{NEW} as a function of $P_{k,m}^t$ and x^1, \dots, x^t .