1 Probability and Bayes Estimation

1. The Binary Symmetric Channel (BSC) can be represented by the transition matrix:
$$P(Y|X) = \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix}.$$
Assuming the marginal probability $p(X=0) = 1-\beta$ and $p(X=1) = \beta$, compute the mutual information $I(X;Y)$.

2. A function $f(x)$ is called convex over (a,b) if for any $x_1, x_2 \in (a,b)$ and $0 \le \lambda \le 1$,
$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2).$$
   (a) Prove that $-\log x$ is a convex function.
   (b) Prove Jensen's inequality:
   $$\mathbb{E}[f(x)] \ge f(\mathbb{E}[x]),$$
   where $f(x)$ is a convex function, and assume the probability distribution is discrete.
   (c) Using Jensen's inequality, prove two properties about the relative entropy $D(P\|Q)$.

3. Consider linear models where the target variable y follows a Gaussian distribution:
$$p(y \mid \mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(y \mid \mathbf{w}^\top \mathbf{x}, \sigma^2).$$
The mean is a linear combination of the features:
$$\mu = \mathbf{w}^\top \mathbf{x} = w_0 + w_1 x_1 + \cdots + w_D x_D.$$
Assuming the parameter $\mathbf{w}$ satisfies the following Gaussian prior distribution:
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \lambda^{-1}\mathbf{I}),$$
   (a) Find the posterior distribution for $\mathbf{w}$, given that the dataset $\mathcal{D}$ consists of N points $\{(y_1, \mathbf{x}_1), \ldots, (y_N, \mathbf{x}_N)\}$.
   (b) Find the loss function from MAP estimation of the posterior probability.

4. Consider a discrete probability distribution with $|A|$ outcomes. Find the probability distribution that maximizes the entropy.