**Mid-term**
ECE 271A
Electrical and Computer Engineering
University of California San Diego

Nuno Vasconcelos Fall 2015

**1.** The entropy of a random variable $X$

$$H[X] = -\int P_X(x) \log P_X(x) dx$$

is a very important quantity in many areas of learning. Unfortunately, the integral above can be intractable for many distributions of practical interest. One frequently used estimator of this quantity, from a sample $\mathcal{D} = \{x_1, \ldots, x_n\}$ of independent observations of $X$, is the sample entropy

$$h[\mathcal{D}] = -\frac{1}{n} \sum_i \log P_X(x_i).$$

In this problem, we study the accuracy of this estimator.

**a) (5 points)** We start by considering the exponential distribution

$$P_X(x; \lambda) = \lambda e^{-\lambda x}, \quad x \geq 0,$$

which is known to have mean $\lambda^{-1}$, variance $\lambda^{-2}$, and entropy $1 - \log(\lambda)$. Show that, for $n$ large enough such that

$$\frac{1}{n} \sum_i X_i \approx E_X[X]$$

holds, the same is true for

$$h[\mathcal{D}] \approx H[X].$$

**b) (10 points)** We next look for a more rigorous way to show this result, applicable to any distribution. Start by showing that for any two *independent* random variables $X$ and $Y$, and any function $f(\cdot)$,

$$E_{X,Y}[\{f(X) - E_X[f(X)]\}\{f(Y) - E_Y[f(Y)]\}] = 0.$$

**c) (10 points)** Compute the mean and variance of the estimates of $h[\mathcal{D}]$. Is this estimator unbiased? What can you say about the asymptotic behavior of the estimator?

**2.** Consider a classification problem with three Gaussian classes

$$P_{\mathbf{X}|Y}(\mathbf{x}|i) = \mathcal{G}(\mathbf{x}, \mu_i, \mathbf{\Sigma}),\ i \in \{1,2,3\},$$

equal class probabilities $P_Y(i) = 1/3, i \in \{1,2,3\}$, means $\mu_i$, and a generic covariance matrix $\mathbf{\Sigma}$, which is equal for all classes. In class, we saw that the Bayes decision rule between classes $i$ and $j$ is a hyperplane. We denote its normal vector by $\mathbf{w}^{ij}$ and the point needed to specify the hyper-plane (in addition to the normal) by $\mathbf{x}_0^{ij}$.

**a)(10 points)** Starting from the Bayes decision rule, derive the expressions of the parameters of hyperplane $ij$ as a function of the parameters of Gaussians $i$ and $j$. *(Note: if you don't know how to derive the expressions, but remember them from class you can simply write them down. However, you will only receive partial credit.)*

**b)(10 points)** Assume that $\mu_1$, $\mu_2$, $\mathbf{w}^{12}$, and $\mathbf{w}^{23}$ are known, and all remaining variables are unknown. Is this sufficient information to determine $\mathbf{w}^{13}$? If so, provide an expression for $\mathbf{w}^{13}$ in terms of the known quantities. If not explain why (including what additional pieces of information would be needed).

**c)(10 points)** Assume that $\mathbf{w}^{12}$, $\mathbf{w}^{23}$, $\mathbf{w}^{13}$ and $\Sigma$ are known, and all remaining variables are unknown. Is this sufficient information to determine $\mathbf{x}_0^{12}$, $\mathbf{x}_0^{23}$, and $\mathbf{x}_0^{13}$? If so, provide an expression for the variables that can be determined, in terms of the known quantities. If not explain why (including what additional pieces of information would be needed).

**3.** It can be shown that a large number of popular probability density functions belong to the *exponential family*. This is the family of distributions of the form

$$P_X(x;\theta) = h(x)\exp\left\{\nu(\theta)T(x) - A(\theta)\right\}.$$

Two well known examples are the Gaussian (of known variance $\sigma^2 = 1$), and binomial densities, which correspond to the choice of functions in the table below.

| pdf | $\theta$ | $\nu(\theta)$ | $A(\theta)$ | $T(x)$ | h(x) |
|---|---|---|---|---|---|
| Gaussian | $\mu$ | $\mu$ | $\frac{\mu^2}{2}$ | $x$ | $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ |
| Binomial | $p$ | $\log\frac{p}{1-p}$ | $-n\log(1-p)$ | $x$ | $\binom{n}{x}$ |

**a) (5 points)** Consider an independent sample $\mathcal{D} = \{x_1, \ldots, x_n\}$ from a random variable $X$ whose distribution is in the exponential family. Show that the maximum likelihood estimate of the parameter $\theta$ must satisfy the equation

$$\frac{\frac{\partial A(\theta)}{\partial\theta}}{\frac{\partial\nu(\theta)}{\partial\theta}} = \frac{1}{n}\sum_{i=1}^{n} T(x_i).$$

Are there any additional requirements?

**b) (10 points)** Consider a classification problem with two classes whose class-conditional distributions are in the same exponential family but have different parameters

$$P_{X|Y}(x|i) = h(x)\exp\left\{\nu(\theta_i)T(x) - A(\theta_i)\right\}, \quad i \in \{0,1\}$$

and class probabilities $P_Y(i) = \pi_i$. A sample of independent measurements $\mathcal{D} = \{x_1, \ldots, x_n\}$ has been collected. It is known that they have all been drawn from the same class, and the goal is to determine that class. Show that the optimal decision function, under the "0/1" loss, for this problem is a threshold on the sample average

$$s_n = \frac{1}{n}\sum_{k=1}^{n} T(x_k).$$

What is this threshold? *(Note that the goal is to classify the entire sample, not one $x_i$ at a time).*

**c) (10 points)** The Kullback-Leibler divergence is a measure between the probability distributions. It is defined as follows

$$KL[P_{X|Y}(x|1)||P_{X|Y}(x|0)] = \int P_{X|Y}(x|1)\log\frac{P_{X|Y}(x|1)}{P_{X|Y}(x|0)}dx.$$

Compute expressions for the KL divergences

$$KL[P_{X|Y}(x|1)||P_{X|Y}(x|0)]$$
$$KL[P_{X|Y}(x|0)||P_{X|Y}(x|1)]$$

as functions of $\nu(\theta_i), A(\theta_i)$, and the expectations $E_{X|Y}[T(x)|Y = i]$.

**d)(10 points)** Using **c)** and the fact that the KL divergence can never be negative, show that the optimal decision rule has zero asymptotic error, i.e. zero error when the sample size $n$ goes to infinity.

**e)(10 points)** After the classifier was built, there was a mistake, and it was fed contaminated samples. Samples from class 0 were perfect, but samples from class 1 were contaminated with points from class

0. More specifically, points in these samples were drawn from class 0 with probability $\epsilon$ ($0 < \epsilon < 1$) and from class 1 with probability $1 - \epsilon$. In summary, rather than being drawn from $P_{X|Y}(X|1)$ the sample from class 1 was drawn from

$$P'_{X|1}(X|1) = \epsilon P_{X|Y}(X|0) + (1 - \epsilon)P_{X|Y}(X|1).$$

Under what conditions does the asymptotic error remain zero?