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

Nuno Vasconcelos

Fall 2018

1. 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

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

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.

$\log p - \log(1$

$+\frac{1}{p}+$

$+\frac{n}{1-p}$

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

Table 1: Examples of exponential family distributions

a) (10 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 θ must satisfy the equation

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

$+\frac{n}{1-p}$

Are there any additional requirements? Use this result to derive the ML estimate of the parameters of the distributions in the table, and check that these satisfy the additional requirements, if there are any.

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

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

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).*

$+\frac{n}{1-p}$

$-\frac{n}{(1-p)^2}$

$\log p - \log(1-p)$

$\frac{1}{p} + \frac{1}{1-p}$ 1

$\left[-\frac{1}{p} - \frac{1}{(1-p)^2}\right] + \frac{\theta^2}{2} = \frac{2\theta}{2}$

$-n$

2. In this problem we consider the invariance of maximum likelihood estimates to model reparameterizations. In all cases we consider an independent and identically distributed binary sample $\mathcal{D} = \{x_1, \ldots, x_n\}$, $x_i \in \{0, 1\} \forall i$.

a) (10 points) Assume that this sample is drawn from a Bernoulli random variable of parameter (probability) θ, i.e.

$$P_X(x; \theta) = \theta^x (1-\theta)^{1-x}. \tag{1}$$

Determine the maximum likelihood estimate θ_{ML} of θ.

b) (10 points) Consider the parameter transformation

$$\gamma = \log\left(\frac{\theta}{1-\theta}\right). \tag{2}$$

Show that, under the new parametrization, X has probability mass function

$$P_X(x; \gamma) = K e^{\gamma x} \tag{3}$$

and determine the constant K.

c) (10 points) Let γ_{ML} be the maximum likelihood estimate of γ. Show that

$$\gamma_{ML} = \log\left(\frac{\theta_{ML}}{1-\theta_{ML}}\right). \tag{4}$$

d) (10 points) Show that the statement above holds for any invertible reparameterization of any model. That is, assume that x has *any* distribution $P_X(x; \theta)$ and

$$\gamma = f(\theta) \tag{5}$$

for some *invertible* transformation $f(.)$. Then show that γ_{ML}, the maximum likelihood parameter estimate under the model $P_X(x; \gamma)$ satisfies the condition

$$\gamma_{ML} = f(\theta_{ML}) \tag{6}$$

where θ_{ML} is the maximum likelihood estimate under the model $P_X(x; \theta)$.

$\frac{\theta}{1-\theta} \times 1$

$\frac{\theta \times 1-\theta}{1-\theta}$

3. Consider a classification problem with two Gaussian classes

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

equal class probabilities $P_Y(i) = 1/2, i \in \{1, 2\}$, means μ_i, and equal covariance Σ. In class, we saw that the Bayes decision rule between these classes is a hyperplane. This hyperplane can be written as

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0.$$

Let $\mathcal{D}_i, i \in \{1, 2\}$ be n-point samples collected from each class. Assume that Σ is known, and μ_i is estimated from sample $\mathcal{D}_i$, through some estimator

$$\hat{\mu}_i = f_i(\mathbf{X}_1, \ldots, \mathbf{X}_n).$$

a) **(10 points)** Starting from the Bayes decision rule, under the "0/1/" loss, derive the expressions of the parameters $\mathbf{w}$ and $\mathbf{x}_0$, as a function of the parameters of the two Gaussians. *(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)** Let $\hat{\mathbf{w}}$ and $\hat{\mathbf{x}}_0$ be "plug-in" estimators of the hyperplane parameters obtained by plugging the estimators $\hat{\mu}_i$ in the formulas derived in **a**). Assuming that Σ is an invertible matrix, can $\hat{\mathbf{w}}$ and $\hat{\mathbf{x}}_0$ both be *unbiased estimators (of* $\mathbf{w}$ *and* $\mathbf{x}_0$*, respectively)* if either $\hat{\mu}_1$ or $\hat{\mu}_2$ is biased? *If not, show why. If yes, what biases are allowed for* $\hat{\mu}_1$ *and* $\hat{\mu}_2$?

c) **(10 points)** Since $\mathbf{w}^T\mathbf{x}_0$ only determines the threshold of the BDR, there are many situations where we don't really care about $\mathbf{x}_0$. In this case, we only need to know if $\hat{\mathbf{w}}$ (as defined in **b**)) is an unbiased estimator of $\mathbf{w}$. Assuming that Σ is an invertible matrix, can this happen if either $\hat{\mu}_1$ or $\hat{\mu}_2$ is biased? *If not, show why. If yes, what biases are allowed for* $\hat{\mu}_1$ *and* $\hat{\mu}_2$?

d) **(10 points)** Repeat **c**) for the case in which $\mathbf{x}$ is d-dimensional, and Σ^{-1} has $d - k$ null eigenvalues. *(Note: You may find it useful to recall that the eigenvector decomposition of a matrix* $\mathbf{A}$ *is,*

$$\mathbf{A} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^T$$

where $\mathbf{\Phi}$ *is the orthonormal matrix whose columns are the eigenvectors of* $\mathbf{A}$*, and* $\mathbf{\Lambda}$ *a diagonal matrix containing the eigenvalues of* $\mathbf{A}$*.)*