Final
ECE 271A
Electrical and Computer Engineering
University of California San Diego

Nuno Vasconcelos

Fall 2013

1. Consider a random variable X such that $P_X(k) = \pi_k, k \in \{1, \ldots, N\}$. Suppose we draw n independent observations from X and form a random vector $\mathbf{C} = (C_1, \ldots, C_N)^T$ where C_k is the number of times that the observed value is k (i.e. $\mathbf{C}$ is the histogram of the sample of observations). Then, $\mathbf{C}$ has multinomial distribution

$$P_{C_1,\ldots,C_N}(c_1, \ldots, c_N) = \frac{n!}{\prod_{k=1}^N c_k!} \prod_{j=1}^N \pi_j^{c_j}.$$

a) (5 points) Derive the ML estimator for the parameters $\pi_i, i = 1, \ldots, N$ from a vector of counts $\mathbf{c}$.

b) (5 points) Derive the MAP estimator of the same parameters from a vector of counts $\mathbf{c}$, under the assumption of a Dirichlet prior

$$P_{\Pi_1,\ldots,\Pi_N}(\pi_1, \ldots, \pi_N) = \frac{\Gamma(\sum_{j=1}^N u_j)}{\prod_{j=1}^N \Gamma(u_j)} \prod_{j=1}^N \pi_j^{u_j - 1}$$

where the u_j are a set of *hyperparameters* (parameters of the prior) and

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$$

the Gamma function.

c) (5 points) Compare the two estimators above. What is the use of the MAP prior equivalent to, in terms of the ML solution? What is the effect of the prior as the number of samples n increases?

2. As discussed in class, one of the practical difficulties of Bayesian estimation is that it is not always clear how to set up prior parameters. In practice, one way to do this is to "cheat," learning the prior parameters from data. There are many techniques to do this, which are broadly known as "empirical Bayes" methods. In this problem we consider two such methods. In all cases, we consider the Gaussian problem that we studied in class, i.e. where we have a scalar Gaussian observation X of unknown mean μ and **known** variance σ^2

$$P_{X|\mu}(x|\mu) = G(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{1}$$

a Gaussian prior of mean μ_0 and variance σ_0^2, $P_\mu(\mu) = G(\mu, \mu_0, \sigma_0^2)$, and an iid sample $\mathcal{D} = \{x_1, \ldots x_n\}$. In class we have seen that, for this problem, the posterior and predictive distributions are Gaussians of mean and variance of the table below, where μ_{ML} is the maximum likelihood estimate of the mean.

	mean	variance	notes
$P_{\mu\|T;\mu_0,\sigma_0^2}(\mu\|\mathcal{D};\mu_0,\sigma_0^2)$	$\mu_n = \alpha_n\mu_{ML} + (1-\alpha_n)\mu_0$	$\sigma_n^2 = \frac{\sigma^2}{n}\alpha_n$	$\alpha_n = \frac{n\sigma_0^2}{\sigma^2+n\sigma_0^2}$
$P_{x\|T;\mu_0,\sigma_0^2}(x\|\mathcal{D};\mu_0,\sigma_0^2)$	μ_n	$\sigma^2 + \sigma_n^2$	

a) the first empirical Bayesian method is equivalent to performing maximum likelihood on the predictive distribution. It sets the prior parameters to

$$(\mu_0^*, (\sigma_0^*)^2) = \arg\max_{\mu_0,\sigma_0^2} P_{x|T;\mu_0,\sigma_0^2}(\mathcal{D}|\mathcal{D};\mu_0,\sigma_0^2). \tag{2}$$

Note that, as usual, $\mathcal{D}$ is the training data. In what follows you only have to check your solution up to the first order optimality conditions, i.e. to show that the desired quantities are critical points of the cost.

i) (5 points) show that $\mu_0^* = \mu_{ML}$.

ii) (10 points) determine $(\sigma_0^*)^2$ as a function of the data in $\mathcal{D}$. Ignore any solution of the form $(\sigma_0^*)^2 = \infty$. (*Hint:* you should build on **i)** here. Plug in μ_0^* in the predictive cost function and simplify as much as possible before solving the optimization)

b) The second method is a true maximum likelihood method. It consists of considering μ a hidden random variable and using EM to determine the ML estimates of the parameters μ_0, σ_0^2

$$(\mu_0^*, (\sigma_0^*)^2) = \arg\max_{\mu_0,\sigma_0^2} P_{x;\mu_0,\sigma_0^2}(\mathcal{D};\mu_0,\sigma_0^2). \tag{3}$$

i) (5 points) derive the log-likelihood $\log P_{x,\mu;\mu_0,\sigma_0^2}(\mathcal{D},\mu;\mu_0,\sigma_0^2)$ of the complete data.

ii) (10 points) Derive the Q function, $Q[(\mu_0,\sigma_0^2)|(\mu_0,\sigma_0^2)^{(i)}]$. What are the computations of the E-step of the algorithm?

iii) (10 points) Determine the parameter updates the M-step of the algorithm.

c) (5 points) We finish by comparing the solutions of the two empirical Bayesian algorithms. What is the value of the parameter estimates to which the EM algorithm converges? By plugging those values on the expression for the predictive distribution, we can obtain that distribution. How does it compare to that based on the optimal parameters found by the predictive-ML method? Under what conditions are the two predictive distributions identical? Derive these conditions for both small and large n.

3. Consider a communications problem involving multiple hops. A signal s_0 is transmitted from an emitter over a link l_1 that adds Gaussian noise to it with known variance σ^2. The signal s_1 received at the end of the link is, therefore, a sample from a Gaussian variable of distribution

$$P_{S_1|S_0}(s_1|s_0) = \mathcal{G}(s_1, s_0, \sigma^2). \tag{4}$$

The signal is then replicated, and sent over a second link (l_2) which, once again, adds Gaussian noise of variance σ^2 to it. This leads to a signal s_2, at the end of link l_2, with distribution

$$P_{S_2|S_1}(s_2|s_1) = \mathcal{G}(s_2, s_1, \sigma^2). \tag{5}$$

The process is repeated over L links, producing a sequence of signals s_i characterized by

$$P_{S_i|S_{i-1}}(s_i|s_{i-1}) = \mathcal{G}(s_i, s_{i-1}, \sigma^2), i \in \{1, \ldots, L\} \tag{6}$$

The original signal s_0 is itself a sample from a Gaussian variable of known mean μ and variance ν^2, i.e.

$$P_{S_0}(s) = \mathcal{G}(s, \mu, \nu^2). \tag{7}$$

The noise added in the different links is not independent, but the system is known to satisfy a Markovian property

$$P_{S_i|S_{i-1},S_{i-2},\ldots,S_0}(s_i|s_{i-1}, s_{i-2}, \ldots, s_0) = P_{S_i|S_{i-1}}(s_i|s_{i-1}). \tag{8}$$

a) (10 points) show that

$$P_{S_i|S_0}(s_i|s_0) = \mathcal{G}(s_i, s_0, i \times \sigma^2), \forall i \in \{1, \ldots, L\} \tag{9}$$

b) (15 points) Suppose that the value of $S_0 = s_0$ is continuously transmitted, leading to the reception of a set of measurements $\mathcal{D}_i = \{x_1^i, \ldots, x_n^i\}$ at the end of link i. Assuming that the x_j^i are sampled independently from $P_{S_i|S_0}(s_i|s_0)$, and that σ^2 is small, answer the following questions.

1. derive the posterior distribution of S_0, given $\mathcal{D}_i, \forall i$.
2. what is the MAP estimate of s_0?
3. consider the cases where i is small ($i \approx 0$) and large ($i \to \infty$). What is the MAP estimate in these extremes? Can you provide an intuitive justification as to why this makes sense?

c) (15 points) Assume that there is a wireless transmitter at the end of each link, and a user, who is walking around, can tune in to the closest transmitter. That is, the user will receive the output of the i^{th} link, S_i, when he is closest to that output. To calibrate the system, the same signal s_0 is continuously transmitted throughout a day. A technician walks around, randomly, throughout the day, and collects a set of measurements $\mathcal{D} = \{x_1, \ldots, x_n\}$. Note that the technician has no knowledge of which output he is closest to, he simply records a signal. Upon returning to his office, he needs to estimate the noise level, σ, of the system.

What are the steps of the algorithm to perform, if he wants to compute the maximum likelihood estimate of σ from the measurements in $\mathcal{D}$? Answer this question, assuming that the technician has equal probability of visiting the L links. Repeat the problem for two cases

1. the transmitted value s_0 is known
2. the transmitted value s_0 is not known