STAT 391

Lecture IV Supplement

Estimation of a uniform distribution

©2007 Marina Meilă
mmp@cs.washington.edu

1 The distribution of the maximum of n points

Here we solve the following probability problem: We have a known distribution over the real line, whose CDF is F. We sample n points $x_1, x_2 \dots x_n$ independently according to F. The maximum of these samples is denoted by y, i.e

$$y \;=\; \max_{i=1:n} x_i \tag{1}$$

Since the points $x_1, x_2 \dots x_n$ are random, y will also vary randomly. The problem is to find the probability distribution of y.

Denote by F_Y the CDF of y. Then, we have

$$\begin{aligned} F_Y(b) = &\quad P[\max_{i=1:n} x_i \le b] \quad \text{(by definition)} && (2)\\ = &\quad P[x_1 \le b \,\text{and}\, \mathrm{x_2} \le \mathrm{a} \,\text{and}\, \dots \mathrm{x_n} \le \mathrm{a}] && (3)\\ = &\quad P[x_1 \le b]P[x_2 \le b] \dots P[x_n \le b] \quad \text{(samples drawn independently)} && (4) \end{aligned}$$

But $P[x_i \le b] = F(b)$ by the definition of the CDF. Therefore

$$F_Y(b) \;=\; F^n(b) \tag{5}$$

This result holds for any distribution.

Example 1 Uniform distribution *If the n data are drawn from a uniform distribution on the interval $[\alpha, \beta]$ then the distribution of the maximum $b = \max\{x_1, x_2 \dots x_n\}$ is given by*

$$F_Y(b) \;=\; \begin{cases} \left(\frac{b-\alpha}{\beta-\alpha}\right)^n, & \text{if } b \in [\alpha, \beta] \\ 0, & \text{if } b < \alpha \\ 1, & \text{if } b > \beta \end{cases} \tag{6}$$

Figure 1 shows this distribution.

Exercise. Prove that the CDF of $Z = \min_{i=1:n} x_i$ is $F_Z(a) = 1 - (1 - F(a))^n$.

1.1 A "confidence" approach to estimating the uniform distribution

Assume now that the data $x_1, x_2 \dots x_n$ are sampled from a uniform distribution $f_{\alpha,\beta}$ with α, β unknown. We have seen that maximum likelihood estimation gives us the values $\alpha^{ML} = \min_{i=1:n} x_i$, $\beta^{ML} = \max_{i=1:n} x_i$ for the parameters α, β and that α^{ML} is always greater α, while β^{ML} is always smaller than the true β.

We shall now derive a method to correct the ML estimates. The method is based on the distribution of the minima and maxima of n samples. Let $\alpha^{ML} = z = \min_{i=1:n} x_i$ and $\beta^{ML} = y = \max_{i=1:n} x_i$. Denote also $l = \beta - \alpha$, $l^{ML} = \beta^{ML} - \alpha^{ML}$ the ranges of the data in the unknown distribution respectively in the observations.

Assuming we know the true parameters, we compute the probability that the data fall in an interval of size l^{ML}

$$
\begin{aligned}
P[\alpha^{ML} \le x_1, x_2, \dots x_n \le \beta^{ML}] &= \prod_{i=1}^{n} P[\alpha^{ML} \le x_i \le \beta^{ML}] && (7)\\
&= \prod_{i=1}^{n} \frac{\beta^{ML} - \alpha^{ML}}{\beta - \alpha} && (8)\\
&= \left(\frac{\beta^{ML} - \alpha^{ML}}{\beta - \alpha}\right)^n && (9)\\
&= \left(\frac{l^{ML}}{l}\right)^n && (10)
\end{aligned}
$$

This probability is plotted in figure 1 (right).

From (10) we get that

$$
l^{ML} = (P[\alpha^{ML} \le x_1, x_2, \dots x_n \le \beta^{ML}])^{1/n} l \qquad (11)
$$

Now we **choose** a value $\gamma = P[\alpha^{ML} \le x_1, x_2, \dots x_n \le \beta^{ML}]$; here γ is a number close to 1, like 95%, and is called **confidence level**. The value $1 - \gamma$ represents the tolerable probability of an error and we set the estimate of $\hat{l}$ to be $\boxed{\hat{l} = l^{ML} / \sqrt[n]{\gamma}}$.

Motivation of this choice is as follows. For any true l, the probability of the observed range l^{ML} to be at most ϵl is ϵ^n. A range l^{ML} that's very small is considered atypical. In figure 1 (right) we see the CDF of the range l^{ML} for the true range $l = 1$. If we cut it at level $1 - \gamma$, and obtain a length $l_{5\%}$ (satisfying $F(l_{5\%}) = 1 - \gamma$), then we call all lengths above this value typical, and all lengths smaller than $l_{5\%}$ atypical. We believe that our observation is typical, hence that it belongs to the top $95\% = \gamma$ of all possible ranges, hence that l^{ML} observed is $\ge l_{5\%}$. Therefore, we can set $P[\alpha^{ML} \le x_1, x_2, \dots x_n \le \beta^{ML} \,|\, l] = \gamma$ in (11) and solve for $\hat{l}$.

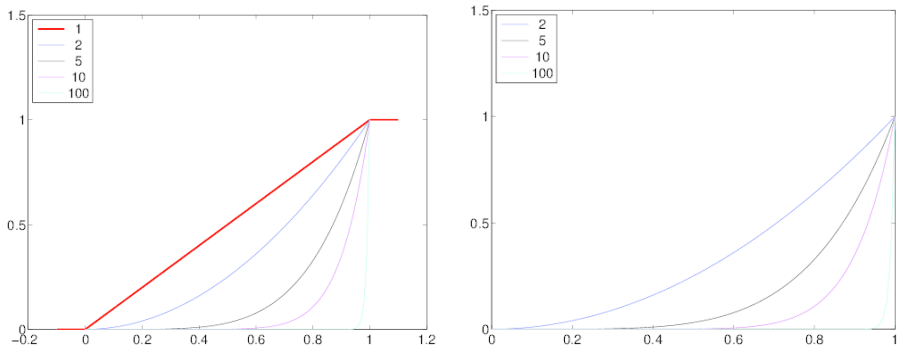

Figure 1: , (Left) The cumulative distribution function (CDF) of the maximum of n samples from a uniform distribution on $[0,1]$ for $n = 1, 2, 5, 10, 100$. (Right) The probability that n samples from a uniform distribution on $[0,1]$ fall in a range of length $l \in (0,1]$ for $n = 2, 5, 10, 100$.

Now we can also understand what the $1-\gamma$ probability of "error" means. The error is to believe that l^{ML} observed is typical at the γ level.

Once $\hat{l}$ is chosen, we correct the interval ends by equal amounts on each side, i.e

$$\hat{\alpha} = \alpha^{ML} - \frac{\beta^{ML} - \alpha^{ML}}{2}(1/\sqrt[n]{\gamma} - 1) \tag{12}$$

$$\hat{\beta} = \beta^{ML} + \frac{\beta^{ML} - \alpha^{ML}}{2}(1/\sqrt[n]{\gamma} - 1) \tag{13}$$

1.2 An alternative correction based on expectations

As we have seen, we always have $\alpha < \alpha^{ML} < \beta^{ML} < \beta$. The correction below insures that the estimates $\hat{\alpha}, \hat{\beta}$ are "centered on" the true values α, β. The motivation and derivations for these formulas is be given in the chapter on expectation.

$$\hat{\alpha} = \alpha^{ML} - \frac{\beta^{ML} - \alpha^{ML}}{n+1} \tag{14}$$

$$\hat{\beta} = \beta^{ML} + \frac{\beta^{ML} - \alpha^{ML}}{n+1} \tag{15}$$

The idea is to estimate the expectation of α^{ML}, β^{ML} as a function of the true parameters.

For simplicity, we take $\alpha = 0, \beta = 1$ since we can obtain the expectations of any other uniform distribution by scaling and shifting the X axis.

In this case, $F_Z(a) = 1 - (1-a)^n$, and therefore the density of the minimum is $f_Z(a) = n(1-a)^{n-1}$. For the maximum we have $F_Y(b) = b^n$ and $f_Y(b) = nb^{n-1}$. We compute the expectations of these distributions.

$$E[Y] = \int_0^1 nb^{n-1}db = \frac{n}{n+1} \tag{16}$$

$$E[Z] = \int_0^1 n(1-a)^{n-1}da = \frac{1}{n+1} \tag{17}$$

As one hoped for, the expectation of the maximum y tends to 1 as n tends to infinity, and that of the minimum tends to 0.

Exercise Show that the variance of Y, Z equals $\frac{n}{(n+1)^2(n+2)}$.

Now to correct the ML estimates, we assume that the observed maximum and minimum α^{ML}, β^{ML} equal their expected values (which close to the truth when the variance is small) and by a simple calculation we obtain the formulas (14).