Probability and mathematical needs

Lab session: solution of the problem

July 9, 2010

4 Problem

Reminder: a Bernouilli distribution with parameter p is a discrete probability distribution 0n $\{0, 1\}$. $X \sim B(p)$ means that $X = 1$ with probability p (and 0 with probability $1 - p$).

Consider the realizations of N i.i.d random variables $X_i \sim B(\theta)$ for $i = 1 \cdots N$, and denote by N_0 the number of 0 obtained, and N_1 the number of 1 obtained. The parameter θ is a random variable, the problem is to estimate the value of θ from the result $N_0,\ N_1$.

1. Maximum likelihood and maximum a posteriori.

   Assume that the prior is $\theta \sim \mathcal{U}[0, 1]$.

   a) Prove that the *maximum a posteriori* θ_{MAP} is also the estimator that maximizes the likelihood of the observation θ_{ML}.

   Let S be the random variable $S = \sum_{i=1}^{N} X_i$. Then in our experiment, we observe $S = N_1$.
   The likelihood is $p_{S|\Theta}(n, \theta) = \theta^n(1-\theta)^{N-n}$ for $n = 0 \cdots N$, $\theta \in [0, 1]$.
   The prior has the density $p_\Theta(\theta) = \delta_{\{\theta \in [0,1]\}}$.
   The posterior is of the form:

$$
\begin{aligned}
p_{\Theta|S}(\theta|n) &= \frac{p_{S|\Theta}(n,\theta) * p_\Theta(\theta)}{\int_{[0,1]} p_{S|\Theta}(n,\theta') * p_\Theta(\theta') d\theta'} \\
&= \frac{p_{S|\Theta}(n,\theta) * 1}{\mathbb{P}_{\{S=n\}}}
\end{aligned}
$$

   Given that we observe $S = N_1$, the maximum likelihood estimator for θ is

$$
\begin{aligned}
\theta_{ML} &= \operatorname{argmax}_{\theta \in [0,1]} p_{S|\Theta}(N_1, \theta) \\
&= \operatorname{argmax}_{\theta \in [0,1]} \theta^{N_1}(1-\theta)^{N-N_1}
\end{aligned}
$$

   and the maximum a posteriori is

$$
\begin{aligned}
\theta_{MAP} &= \operatorname{argmax}_{\theta \in [0,1]} p_{\Theta|S}(\theta, N_1) \\
&= \operatorname{argmax}_{\theta \in [0,1]} \frac{p_{S|\Theta}(n,\theta) * 1}{\mathbb{P}\{S=n\}} \\
&= \operatorname{argmax}_{\theta \in [0,1]} p_{S|\Theta}(n, \theta)
\end{aligned}
$$

   Thus $\theta_{MAP} = \theta_{ML}$.

b) Compute θ_{ML}.

Given that we observe $S = N_1$, the maximum likelihood estimator for θ is

$$\begin{aligned}\theta_{ML} &= \text{argmax}_{\theta\in[0,1]}\, p_{S|\Theta}(N_1, \theta) \\ &= \text{argmax}_{\theta\in[0,1]}\, \theta^{N_1}(1-\theta)^{N-N_1} \\ &= \text{argmax}_{\theta\in[0,1]}\, e^{N_1\log(\theta)+(N-N_1)\log(1-\theta)} \\ &= \text{argmax}_{\theta\in[0,1]}\, N_1\log(\theta) + (N-N_1)\log(1-\theta)\end{aligned}$$

The function $g : \theta \to N_1\log(\theta) + (N-N_1)\log(1-\theta)$ is concave and differentiable. Its derviative is: $g'(\theta) = \frac{N_1}{\theta} - \frac{N-N_1}{1-\theta}$ hence $g' = 0$ yields: $\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_0+N_1}$

2. Maximum a posteriori

Assume that the prior is a truncated Gaussian: $p_\theta(x) = \frac{1}{Z}\exp\left\{-\frac{x^2}{2s}\right\}\delta_{[0,1](x)}$ (where Z is a normalizing constant). I.e θ is distributed as a Gaussian with variance s on $[0,1]$ but never takes values outside of $[0,1]$. We will write $\theta \sim \mathcal{N}(0,s)|_{[0,1]}$.

a) Prove that the *maximum a posteriori* θ_{MAP} maximizes a function f that is a sum of the log-likelihood of the observation plus a quadratic term. (The quadratic term is called the regularization term.)

The prior has the density $p_\Theta(\theta) = \frac{1}{Z}\delta_{\{\theta\in[0,1]\}}e^{-\frac{\theta^2}{2s}}$.
The posterior is of the form:

$$\begin{aligned}p_{\Theta|S}(\theta|n) &= \frac{p_{S|\Theta}(n,\theta) * p_\Theta(\theta)}{\mathbb{P}S=n} \\ &= \theta^{N_1}(1-\theta)^{N-N_1}\frac{1}{Z}\delta_{\{\theta\in[0,1]\}}e^{-\frac{\theta^2}{2s}}\end{aligned}$$

Given that we observe $S = N_1$, the maximum a posteriori is

$$\begin{aligned}\theta_{MAP} &= \text{argmax}_\theta\, \theta^{N_1}(1-\theta)^{N-N_1}\frac{1}{Z}\delta_{\{\theta\in[0,1]\}}e^{-\frac{\theta^2}{2s}} \\ &= \text{argmax}_{\theta\in[0,1]}\, e^{N_1\log(\theta)+(N-N_1)\log(1-\theta)}e^{-\frac{\theta^2}{2s}} \\ &= \text{argmax}_{\theta\in[0,1]}\, N_1\log(\theta) + (N-N_1)\log(1-\theta) - \frac{\theta^2}{2s}\end{aligned}$$

Consider the function $f : \theta \in [0,1] \to N_1\log(\theta) + (N-N_1)\log(1-\theta) - \frac{\theta^2}{2s}$. We have $f = g - \frac{\theta^2}{2s}$ where g is the log-likelihood we have seen in the previous question, and $\theta \to -\frac{\theta^2}{2s}$ is the quadratic term.

b) Prove that the regularized maximization problem (maximizing f) is still a concave problem. Prove that the maximum θ_{MAP} is the zero of a third degree polynomial.

f is twice differentiable. $f'(\theta) = \frac{N_1}{\theta} - \frac{N-N_1}{1-\theta} - \frac{\theta}{s^2}$ and $f''(\theta) = -\frac{N_1}{\theta^2} - \frac{N-N_1}{(1-\theta)^2} - \frac{\theta}{s} \le 0$ hence f is concave. θ_{MAP} is the point where $f' = 0$: it is the point which verifies:

$$\begin{aligned}f'(\theta) = 0 &\Leftrightarrow \frac{N_1}{\theta} - \frac{N-N_1}{1-\theta} - \frac{\theta}{s^2} = 0 \\ &\Leftrightarrow \frac{N_1(1-\theta)-(N-N_1)\theta}{\theta(1-\theta)} = \frac{\theta}{s^2} \\ &\Leftrightarrow s^2(N_1(1-\theta) - (N-N_1)\theta) = \theta^2(1-\theta) \\ &\Leftrightarrow \theta^3 - \theta^2 - s^2N\theta + s^2N_1 = 0\end{aligned}$$

Hence θ_{MAP} is a zero of the above third degree polynomial.

3. Experimentation

   Assume that the prior is still a truncated Gaussian $\theta \sim \mathcal{N}(0, s)|_{[0,1]}$. The goal is to compare numerically the maximum likelihood and the maximum a posteriori.

   a) Create an Octave script *lab.m* that simulates the previous problem:
      - Generate the true parameter θ_{true} from its law $\mathcal{N}(0, s)|_{[0,1]}$ with `s=0.1`.
      - Generate $N = 20$ realizations of $X_i \sim B(\theta_{true})$. Let N_0 (resp. N_1) be the number of 0 (resp. 1) obtained.
      - Compute the estimator that maximizes the likelihood of the observation θ_{ML}.

      (use `normrnd` to generate a realization of a Gaussian, `unifrnd` to generate a uniform).
   b) Using the function `roots`, compute the *maximum a posteriori* θ_{MAP}.
   c) Repeat the experiment. Change the value of the standard deviation `s`. Can one say that θ_{MAP} is closer from θ_{true} than θ_{ML}?

```
% Generate the true parameter
s=0.1;
p=normrnd(1/2,s);
while (p<0 || p>1)
p=normrnd(1/2,s)
end

% Generate N = 20 realizations of X_i ~ B(θ_true).  Let N_0 (resp.  N_1)
% be the number of 0 (resp.  1) obtained.
x=unifrnd(1,20);
N1=length(find(x<p));
N0= length(find(x>=p));

%Compute θ_ML.
disp(N1/(N0+N1))

% Compute the θ_MAP
disp(roots([1,-1,-s(N0+N1),s(N1)]))
```

4. Gradient ascent

   To obtain an estimate of θ_{MAP}, one can perform a gradient ascent on f.

   Reminder: A gradient descent is an iterative algorithm of the form:

$$x^{t+1} = x^t - \gamma_t \nabla f(x^t).$$

   Here we implement the ascent with a fixed step $\gamma_t = \gamma > 0$.

   a) Create two inline functions `f` and `f1=`∇`f`. `> f = inline(' N1*log(x)+ N0*log(1-x)-x/2/s`
      `> f1 = inline('N1./x- N0./(1-x)-x/s','x','N0','N1','s')`
   b) Create a function `x = mygradientascent(f,f1,x0,`γ`,n)` that take as arguments:
      - `f` the inline function to maximize, and `f1` its gradient (defined inline as well),

- `x0` the initial point,
- $\gamma > 0$ the fixed stepsize,
- `n` the number of iteration,

and outputs the n-th iterate of the gradient ascent of `f` starting from `x0`.

c) Experiment with different values of `n`, γ, `x0`. Do you always find θ_{MAP} ? Why ?

```
 function x = mygradientascent(f,f1,x0,γ,n)
% x = mygradientascent(f,f1,x0,γ,n)
x= x0;
for i=1:n
   x = x + γ*f1(x);
end
```