Task 1

Monday, 5 November 2018 16:01

1 Optimizing Likelihoods: Monotonic Transforms

Usually one considers the *log-likelihood*, $\log p(x_1, \ldots, x_n \mid \theta)$. The next problems justify this.

In the lecture, we encountered the likelihood maximization problem

$$\arg\max_{\theta \in [0,1]} \theta^t (1-\theta)^h,$$

where t and h denoted the number of tails and heads in a sequence of coin tosses, respectively.

Problem 1: Compute the first and second derivative of this likelihood w.r.t. θ. Then compute first and second derivative of the log likelihood $\log \theta^t (1-\theta)^h$.

Naïve approach : working with likelihood function:

$$f(\theta) = \theta^t (1-\theta)^h$$

$$\frac{\partial}{\partial\theta} f(\theta) = t\theta^{t-1}(1-\theta)^h - h\theta^t (1-\theta)^{h-1}$$

$$= \theta^{t-1}(1-\theta)^{h-1}\left(t(1-\theta) - h\theta\right)$$

$$\frac{\partial}{\partial\theta^2} f(\theta) = \theta^{t-2}(1-\theta)^{h-2}\left[(t-1)(1-\theta) - (h-1)\theta\right]$$
$$+ \theta^{t-1}(1-\theta)^{h-1}\left[-t-h\right]$$

Working with $f(\theta)$ (likelihood is hard) $\Rightarrow$ log. likelihood

log. likelihood $= \log f(\theta) = g(\theta)$

$$g(\theta) = \log\left[\theta^t (1-\theta)^h\right] = t\log(\theta) + h\log(1-\theta)$$

$$\frac{\partial}{\partial\theta} g(\theta) = t\theta^{-1} - h(1-\theta)^{-1}$$

$$\frac{\partial^2}{\partial\theta^2} g(\theta) = -t\,\theta^{-2} - h(1-\theta)^{-2}$$

Problem 2: Show that every local maximum of $\log f(\theta)$ is also a local maximum of the differentiable, positive function $f(\theta)$. Considering this and the previous exercise, what is your conclusion?

$\{\text{local maximas of } \log f(\theta)\} = \{\text{local maximas of } f(\theta)\}$

The proof

① θ_* is a local maxima of $g(\cdot)$
$\Rightarrow$ θ_* is a local maxima of $f(\cdot)$

② θ_* is a local maxima of $f(\cdot)$
θ_* is a local maxima of $g(\cdot)$

Part ① $\theta \in [\theta_* - \xi, \theta_* + \xi]$ $\quad g(\theta_*) \geq g(\theta)$ $\qquad g(\theta) = \log f(\theta)$

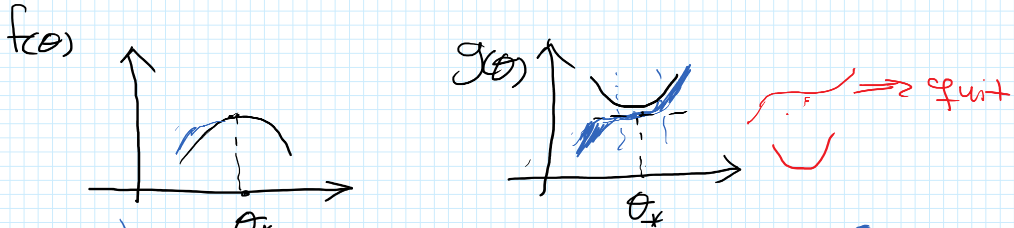
$\Rightarrow \exp(g(\theta_*)) \geq \exp(g(\theta)) \Rightarrow f(\theta_*) \geq f(\theta)$

$\Rightarrow$ θ_* a local maxima for f.

Part ② θ_* is a local maxima of f

$$\Rightarrow \frac{\partial}{\partial\theta} f\Big|_{\theta=\theta_*} = 0 \Rightarrow \frac{\partial}{\partial\theta} g(\theta)\Big|_{\theta_*} = \frac{1}{f(\theta)} \frac{\partial}{\partial\theta} f(\theta)\Big|_{\theta_*} = 0$$

$\Rightarrow$ θ_* is either a local maxima or a local minima



like Part 1, we can show $\Rightarrow$ θ_* is a local maxima of g

∂^2

like last 1, we can show maxima of g

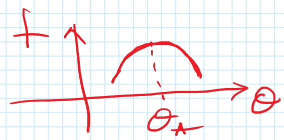
$$\frac{\partial^2}{\partial\theta^2} g(\theta) < 0$$

$$\frac{\partial}{\partial\theta} g(\theta) = f(\theta)^{-1} f'(\theta)$$

$$\frac{\partial^2}{\partial\theta^2} g(\theta) = (-1)\, f(\theta)^{-2} f'(\theta)^2 + f(\theta)^{-1} f''(\theta)$$

$$= f(\theta)^{-1}\left(- f(\theta)^{-1} \underbrace{f'(\theta)^2}_{0} + f''(\theta)\right)\Big|_{\theta=\theta_*}$$

$$= \underbrace{f(\theta_*)^{-1}}_{>0}\ \underbrace{f''(\theta_*)} < 0$$



Task 2

Monday, 5 November 2018 16:01

Problem 3: Show that θ_{MLE} can be interpreted as a special case of θ_{MAP} in the sense that there always exists a prior $p(\theta)$ such that $\theta_{\text{MLE}} = \theta_{\text{MAP}}$.

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\ p(D|\theta)$$

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\ p(D|\theta)\, p(\theta)$$

when $p(\theta)$ is independent of θ

$\Rightarrow \theta \quad = \theta_{MAP}$

$p([\theta_1, \dots, \theta_k])$

$= p(\theta_1) \cdots p(\theta_k)$

where $p(\theta_i) \sim \text{Uniform}(\cdot)$.

Coin Example

$P(\theta) = \text{Beta}(\theta; a, b)$

$$\theta_{MAP} = \frac{a + \#[T] - 1}{a + b + T + H - 2}$$

$a = 1,\ b = 1$

$$\Rightarrow \theta_{MAP} = \frac{T}{T + H}$$

How can we make $p(\theta)$ indepent of θ?

- if domain of θ is finite $\Rightarrow$ $P(\theta) = \text{Uniform}$
- Otherwise? e.g. $\theta \in \mathbb{R}$? Improper Priors.

Problem 4: Consider a Bernoulli random variable X and suppose we have observed m occurrences of $X = 1$ and l occurrences of $X = 0$ in a sequence of $N = m + l$ Bernoulli experiments. We are only interested in the number of occurrences of $X = 1$—we will model this with a Binomial distribution with parameter θ. A prior distribution for θ is given by the Beta distribution with parameters a, b. Show that the posterior *mean* value $\mathbb{E}[\theta \mid \mathcal{D}]$ (not the MAP estimate) of θ lies between the prior mean of θ and the maximum likelihood estimate for θ.

To do this, show that the posterior mean can be written as λ times the prior mean plus $(1-\lambda)$ times the maximum likelihood estimate, with $0 \le \lambda \le 1$. This illustrates the concept of the posterior mean being a compromise between the prior distribution and the maximum likelihood solution.

The probability mass function of the Binomial distribution for some $m \in \{0, 1, \ldots, N\}$ is

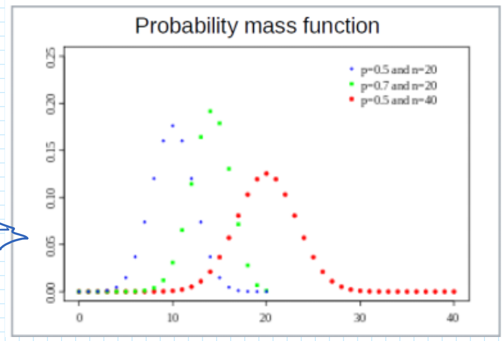
$$p(x = m \mid N, \theta) = \binom{N}{m} \theta^m (1-\theta)^{N-m}.$$

Hint: Identify the posterior distribution. You may then look up the mean rather than computing it.

$X =$ number of Tails when tossing a coin N times.

$$P(X=n) = \binom{N}{n} \theta^{n} (1-\theta)^{N-n}$$

Binomial distribution



Notation	$B(n, p)$
Parameters	$n \in \mathbf{N}_0$ — number of trials $p \in [0,1]$ — success probability in each trial
Support	$k \in \{0, \ldots, n\}$ — number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$

$n \in \{0, 1, 2, \ldots, N\}$

Observations $= D = \{m_1, m_2, \ldots, m_M\}$

$m_j \in \{0, \ldots, N\}$

For simplicity we assume that $D = \{m\}$

$$P(\theta \mid D) \propto P(D \mid \theta) \; P(\theta)$$

$$\left[\binom{N}{m} \theta^{m} (1-\theta)^{N-m}\right] \left[\text{const of Beta}(a,b) \; \theta^{a-1} (1-\theta)^{b-1}\right]$$

$$\propto \theta^{m+a-1} (1-\theta)^{N+b-m-1}$$

$= \text{Beta}(\theta; \underline{\underline{a}}+m, \underline{\underline{b}}+N-m)$

In general $\mathop{E}_{x \sim \text{Beta}(\alpha,\beta)}[x] = \dfrac{\alpha}{\alpha+\beta}$

$$\mathop{E}_{\theta \sim p(\theta|D)}[\theta] = \frac{a+m}{a+\cancel{m}+b+N-\cancel{m}} = \frac{a+m}{a+b+N}$$

$$= \underbrace{\frac{\cancel{a+b}}{a+b+N}}_{(1-\lambda)} \times \underbrace{\frac{a}{\cancel{a+b}}}_{\text{Exp. of prior}} + \underbrace{\frac{\cancel{N}}{a+b+N}}_{\lambda} \times \underbrace{\frac{m}{\cancel{N}}}_{\theta_{MLE}}$$

$$\Rightarrow \underbrace{\mathop{E}_{\theta \sim p(\theta|D)}[\theta]}_{} = (1-\lambda) \mathop{E}_{\theta \sim p(\theta)}[\theta] + \lambda \, \theta_{MLE}$$

---

x_1 . . . . .

$p(\theta|D) \sim \text{Beta}$ -

$$p(x_{N+1} | x_1, \ldots, x_N) = \mathop{E}_{\theta \sim \text{Posterior } p(\theta|D)}[p(x_{N+1}|\theta)]$$

Problem 5:

(a) The definition of an unbiased estimator is as follows: Let X be a random variable with probability density function $p(X|\lambda)$. Let $\{X_1, ..., X_n\}$ be n i.i.d. samples from X. An estimator λ_{EST} for λ is called unbiased iff

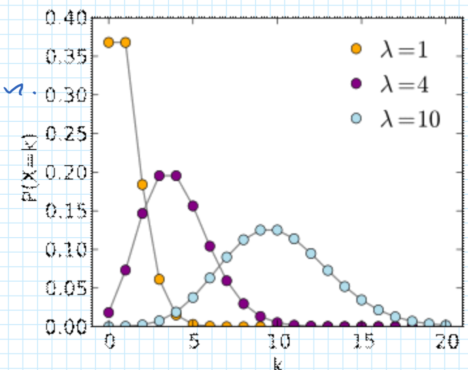
$$\mathbb{E}\Big[\lambda_{EST}(X_1, ..., X_n)\Big] = \lambda. \tag{1}$$

Note that, as denoted in the above equation, the estimator λ_{EST} is a function of the samples.

Let X be Poisson distributed. For n i.i.d. samples from X, determine the maximum likelihood estimate for λ. Show that this estimate is unbiased!

(b) In class we also talked about avoiding overfitting of parameters via *prior* information. Compute the posterior distribution over λ, assuming a Gamma(α, β) prior for it. Compute the MAP for λ under this prior. Show your work.

MLE estimate for λ of a Poisson distribution.



$k \in \{0, 1, 2, ...\}$

$$\text{pdf} = f(x; \lambda) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

Observations $= D = \{x_1, ..., x_N\}$

$$\text{likelihood} = \prod_{n=1}^{N} P(x_n | \lambda)$$

$$= \prod_{n=1}^{N} \left[\frac{e^{-\lambda} \lambda^{x_n}}{x_n!} \right] = \frac{e^{-\lambda N} \lambda^{\sum x_n}}{x_1! \cdots x_N!}$$

$$\Rightarrow \text{log likelihood} = \log P(D|\lambda) = -\lambda N + \left(\sum x_n\right) \log(\lambda)$$
$$- \log(x_1! \cdots x_N!)$$

$$\frac{\partial}{\partial \lambda} \log P(D|\lambda) \overset{!}{=} 0 \Rightarrow -N + \frac{\sum x_n}{\lambda} \overset{!}{=} 0$$

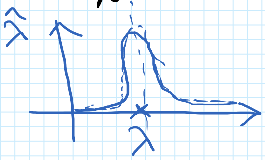
$$\Rightarrow \lambda = \frac{\sum x_n}{\quad} \qquad \frac{\partial^2}{\partial \lambda^2} \boxed{} = (-1) \lambda^{-2} \left(\sum x_n\right) < 0$$

$$\Rightarrow \boxed{\lambda_{MLE} = \frac{\sum x_n}{N}} \qquad \frac{\partial^2}{\partial \lambda^2}\,\boxed{} = (-1)\,\lambda^{-2}\left(\sum x_n\right) < 0$$

Why it is unbiased?

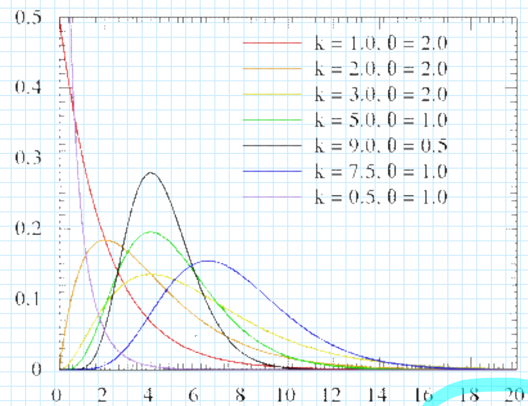
$$E[\lambda_{MLE}] = E\left[\frac{\sum x_n}{N}\right] = \frac{1}{N}\sum_n \overbrace{E[x_n]}^{\lambda} = \frac{N\lambda}{N} = \lambda$$

$$\hat{\lambda}(x_1, \ldots, x_N) \sim$$



$$S^2 = \frac{1}{N-1}\sum (x_i - \bar{x})^2$$

$$P(\lambda) = Gamma(\lambda; \alpha, b)$$



$$P(\lambda|D) \propto P(D|\lambda)\;P(\lambda)$$

$$\left[\frac{e^{-\lambda N}\;\lambda^{\sum x_i}}{x_1!\ldots x_N!}\right]\left[const(\alpha, b)\;e^{-\beta\lambda}\;\lambda^{\alpha-1}\right]$$

$$\propto e^{-\lambda(N+\beta)}\;\lambda^{\sum x_i + \alpha - 1}$$

pdf: $const(\alpha, b) \times e^{-\beta x}\;x^{\alpha-1}$

$$= Gamma\left(\lambda;\; \alpha + \sum x_i,\; N+\beta\right)$$

Now let's assume $P(\lambda) = Beta(\lambda; \alpha, \beta)$

$$P(\theta|D) \propto P(D|\theta)\;P(\theta)$$

$P(\theta|D) = P(D|\theta)P(\theta)\;\boxed{}$

$\Rightarrow \log P(\theta|D) = \log P(D|\theta) + \log P(\theta) + \boxed{}$

$$\left[\frac{e^{-\lambda N}\;\lambda^{\sum x_i}}{x_1!\ldots x_N!}\right]\left[const(\alpha, \beta)\;\lambda^{\alpha-1}(1-\lambda)^{\beta-1}\right]$$

$$\propto e^{-\lambda N}\;\lambda^{\alpha + \sum x_i - 1}\;(1-\lambda)^{\beta-1}$$

Not Beta Distrib.

Not Gamma Distrib.

Compute θ_{MAP} Posterior $= P(\theta|D) = \text{Gamma}(\theta; \alpha + \sum x_i, \beta + N)$

$$\Rightarrow \theta_{MAP} = \frac{\alpha + \sum x_i - 1}{\beta + N}$$

What if we choose $P(\lambda) = \text{Poisson}(\lambda; \alpha)$

$$\Rightarrow P(\lambda|D) \propto P(D|\lambda) \; P(\lambda)$$

$$\left[\frac{e^{-\lambda N} \quad \lambda^{\sum x_i}}{x_1! \cdots x_N!} \right] \left[\frac{e^{-\alpha} \alpha^{\lambda}}{\lambda!} \right]$$

$$\propto e^{-\lambda N} \lambda^{\sum x_i} \frac{\alpha^{\lambda}}{\lambda!} \leftarrow$$

The other reason is that $\lambda \in (0, +\infty)$

But the support of Poisson is $\{0, 1, 2, \cdots\}$

--- Posterior $\propto e^{-\lambda(N+\beta)} \lambda^{\sum x_i + \alpha - 1}$

$$= e^{-\lambda(N+\beta)} \lambda^{\sum x_i + \alpha - 1} \left[\int e^{-\lambda(N+\beta)} \lambda^{\sum x_i + \alpha - 1} \, d\lambda \right]^{-1}$$

$$= e \quad \lambda \quad \left[\int e \quad \lambda \quad d\lambda \right]$$

$$\text{Gamma}(x; \alpha, \beta) = \text{const}(\alpha, \beta)\, x^{\alpha-1} e^{-\beta x}$$

$$\int \text{Gamma}(x; \alpha, \beta)\, dx = 1 \Rightarrow \underbrace{\int x^{\alpha-1} e^{-\beta x}\, dx}\, = \frac{1}{\text{const}(\alpha, \beta)}$$

$$\text{Posterior} = e^{-\lambda(N+\beta)} \lambda^{\sum x_i + \alpha - 1} \times \text{const}\left(\sum x_i + \alpha, N + \beta\right)$$

$$= \text{Gamma}\left(\sum x_i + \alpha, N + \beta\right)$$