Likelihood and Maximum Likelihood Estimation

Huizhen Yu

janey.yu@cs.helsinki.fi
Dept. Computer Science, Univ. of Helsinki

Probabilistic Models, Spring, 2010

Outline

Outline

Definitions

Let y be the observed value of a discrete random variable Y.

Consider a model Θ with each $\theta \in \Theta$ specifying a probability distribution of Y. We denote this distribution by $P(Y;\theta)$ and its PMF by $p(y;\theta)$.

- *Likelihood* for θ based on y:

$$L(\theta) = P(Y = y; \theta), \quad \theta \in \Theta.$$

  It is a function of θ for fixed y. (For fixed θ, $L(\theta)$ is a random variable.)
- *Log likelihood*: $\ell(\theta) = \ln L(\theta)$.
- To emphasize the depedence of the likelihood on data, it can help to write $L(\theta; y)$ and $\ell(\theta; y)$.
- Likelihood is a natural basis for assessing the plaussibility of θ. Maximum likelihood estimation:

$$\max_{\theta \in \Theta} \ell(\theta; y) \quad \textit{or} \quad \max_{\theta \in \Theta} L(\theta; y).$$

  The parameter $\hat{\theta}$ that maximizes the likelihood function is called the *maximum likelihood estimate* of θ.

Invariance Properties of Likelihood Function

Likelihood is invariant to one-one reparametrization of parameters:

- Suppose our model is parametrized by ψ and $\theta = \theta(\psi)$ is a one-one transformation of ψ. Then $P(Y; \psi) = P(Y; \theta(\psi))$, so

$$L(\psi; y) = L(\theta(\psi); y).$$

- This shows that we can choose a suitable parametrization for a particular problem.

Likelihood is invariant to known one-one transformations of the data:

- Suppose Z is a known one-one transformation of Y.
  - If Y is a discrete random variable, $L(\theta; y) = L(\theta; z)$ certainly.
  - If Y is a continuous random variable and $Z = Z(Y)$ is a differentiable one-one transformation of Y, then the density of Z is

$$f_Z(z; \theta) = f_Y(y; \theta)\, |dy/dz|$$

    where $|dy/dz|$ is the determinant of the Jacobian matrix of the transformation from Z to Y. So

$$\ell(\theta; z) = \ell(\theta; y) + \text{ some constant.}$$

  - This shows that in the continuous case, within a particular model, the absolute value of the likelihood is irrelevant to inference about θ.

Example I

Terminology: *a random sample of size n* refers to a collection of n independent, identically distributed random variables.

Suppose we observe the values $y = (y_1, y_2, \ldots, y_n)$ of a random sample of size n. Our model is

$$\Theta = \Big\{ (\theta_1, \ldots, \theta_m) \,\Big|\, \sum_i \theta_i = 1,\ \theta_i \geq 0,\ \forall i \Big\},$$

$$\text{and} \qquad P(Y_k = i; \theta) = \theta_i, \quad i = 1, \ldots, m.$$

Then

$$L(\theta; y) = \prod_{k=1}^{n} p(y_k; \theta) = \prod_{i=1}^{m} \theta_i^{n_i},$$

$$\ell(\theta; y) = \sum_{k=1}^{n} \ln p(y_k; \theta) = \sum_{i=1}^{m} n_i \ln \theta_i,$$

where n_i is the number of occurences of i in y.

Example I

To maximize the likelihood, we solve $\max_{\theta \in \Theta} L(\theta; y)$, or equivalently,

$$\max_{\theta \in \Theta} \ \ell(\theta; y) = \max_{\theta \in \Theta} \Big\{ \sum_{i=1}^{m} n_i \ln \theta_i \Big\}. \qquad (1)$$

The solution is, (to be shown shortly),

$$\hat{\theta}_i = \frac{n_i}{n_1 + n_2 + \cdots + n_m}, \quad i = 1, 2, \ldots, m.$$

I.e., the maximum likelihood estimate $\hat{\theta}$ coincides with the observed frequencies of $1, 2, \ldots, m$ in y.

Maximization problems of the form (1) appear often in maximum likelihood estimation. We show another example next, and we then prove the above statement using the information inequality.

Example II: Fitting a Markov Model

Suppose under model Θ, $Y = (Y_1, Y_2, \ldots, Y_n)$ is a homogeneous Markov chain on $S = \{1, 2, \ldots, m\}$.

The parameter $\theta \in \Theta$ consists of the initial distribution, denoted by $\mu_i, i \in S$, and the set of transition probabilities, denoted by $\{\theta_{ij}, i \in S, j \in S\}$.

We observe $Y = y = (y_1, y_2, \ldots, y_n)$. Then,

$$L(\theta; y) = p(y_1; \theta) \prod_{k=2}^{n} p(y_k \,|\, y_{k-1}; \theta) = \mu_{y_1} \prod_{i \in S} \prod_{j \in S} \theta_{ij}^{n_{ij}}$$

where n_{ij} is the number of transitions from state i to j observed in the sequence y. And

$$\ell(\theta; y) = \ln p(y_1; \theta) + \sum_{k=2}^{n} \ln p(y_k \,|\, y_{k-1}; \theta)$$
$$= \ln \mu_{y_1} + \sum_{i \in S} \sum_{j \in S} n_{ij} \ln \theta_{ij}.$$

Example II: Fitting a Markov Model

Let $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{im})$ denote the vector of transition probabilities from state i, and let Δ denote the space of θ_i:

$$\Delta = \Big\{(z_1, z_2, \ldots, z_m) \,\Big|\, \sum_{j \in S} z_j = 1,\ z_j \geq 0,\ j \in S\Big\}.$$

Maximizing the log likelihood is equivalent to m separate maximization problems:

$$\max_{\theta \in \Theta} \ell(\theta; y) \ \Leftrightarrow\ \max_{\theta_1 \in \Delta} \max_{\theta_2 \in \Delta} \cdots \max_{\theta_m \in \Delta} \Big\{ \sum_{i \in S} \sum_{j \in S} n_{ij} \ln \theta_{ij} \Big\} = \sum_{i \in S} \max_{\theta_i \in \Delta} \Big\{ \sum_{j \in S} n_{ij} \ln \theta_{ij} \Big\}.$$

The solution $\hat{\theta}_i$ of each subproblem $\max_{\theta_i \in \Delta} \Big\{ \sum_{j \in S} n_{ij} \ln \theta_{ij} \Big\}$ is

$$\hat{\theta}_{ij} = \frac{n_{ij}}{n_{i1} + n_{i2} + \cdots + n_{im}}, \quad j \in S.$$

I.e., the maximum likelihood estimates $\hat{\theta}_i, i \in S$ coincide with the observed frequencies of transitions in the sequence y.

Note that each subproblem has the same form as the problem in (1).

Outline

Information Inequality

Let $p = (p_1, p_2, \ldots, p_m), q = (q_1, q_2, \ldots, q_m)$ be non-negative vectors in $\Re^m$. Then,

$$\sum_i q_i \ln p_i \leq \sum_i q_i \ln q_i + \sum_i p_i - \sum_i q_i, \tag{2}$$

with equality if and only if (iff.) $p = q$. (We define $0 \cdot (-\infty) = 0$ in the above.)

When p, q correspond to PMFs, $\sum_i p_i = \sum_i q_i = 1$, and inequality (2) simplifies to

$$\sum_i q_i \ln p_i \leq \sum_i q_i \ln q_i \tag{3}$$

with equality iff. $p = q$.

The difference between the right and left-hand sides of (3) is indeed the *Kullback-Leibler divergence* between q and p:

$$\mathrm{KL}(q, p) = \sum_i q_i \ln(q_i / p_i).$$

Information Inequality

Let X be a discrete random variable with distribution Q and PMF q.
Let p be the PMF of another probability distribution P on the space of X.
Let $q_i = q(i) = Q(X = i),\ p_i = p(i) = P(X = i),\ \forall i$.

- Entropy of X:

$$H(X) = -\sum_i q_i \ln q_i = \mathrm{E}_Q\big[-\ln q(X)\big]$$

- KL-divergence between q and p:

$$\mathrm{KL}(q, p) = \sum_i q_i \ln(q_i / p_i) = \mathrm{E}_Q\Big[\ln\Big(\frac{q(X)}{p(X)}\Big)\Big].$$

($\mathrm{E}_Q[\cdots]$ denotes expectation over X with respect to Q.)

Inequality (3), $\sum_i q_i \ln p_i \leq \sum_i q_i \ln q_i$ ("=" iff. $p = q$), is identical to

$$\mathrm{KL}(q, p) \geq 0 \quad \textit{and} \quad \mathrm{KL}(q, p) = 0 \ \text{ iff. } p = q.$$

Derivation of Inequality (2)

Consider the function $\ln x$ on $x > 0$.

A first-order approximation of $\ln x$ at any $\bar{x} > 0$ always lies above $\ln x$:

$$\ln x \le \ln \bar{x} + \frac{1}{\bar{x}}(x - \bar{x}), \quad \forall\, x, \bar{x} > 0, \tag{4}$$

and equality holds iff. $x = \bar{x}$.

Multiplying both sides by $\bar{x}$, we have

$$\bar{x} \ln x \le \bar{x} \ln \bar{x} + x - \bar{x}, \quad \forall\, x, \bar{x} \ge 0, \tag{5}$$

with equality iff. $x = \bar{x}$. In the above we have also extended the inequality to include the case $x = 0$ or $\bar{x} = 0$, and we define $0 \cdot (-\infty) = 0$.

Applying (5) to bound $q_i \ln p_i$ with $\bar{x} = q_i, x = p_i$,

$$q_i \ln p_i \le \left(q_i \ln q_i + p_i - q_i\right), \quad \Rightarrow \quad \sum_i q_i \ln p_i \le \sum_i q_i \ln q_i + \sum_i p_i - \sum_i q_i,$$

and equality holds iff. $p = q$. This establishes the information inequality (2).

Implications of Information Inequality

Let X, Y be discrete random variables with joint distribution P.

- The *mutual information* between X and Y is defined as

$$I(X;Y) = \mathrm{E}\left[\ln\left(\frac{p(X,Y)}{p(X)p(Y)}\right)\right],$$

  and equivalently,

$$I(X;Y) = \sum_{x,y} p(x,y) \ln\left(\frac{p(x,y)}{p(x)p(y)}\right).$$

- By the information inequality,

$$I(X;Y) \ge 0, \quad \text{and} \quad I(X;Y) = 0 \ \text{ iff. } \ X \perp Y.$$

Implications of Information Inequality

Let X, Y, Z be discrete random variables with joint distribution P.

- The *conditional mutual information* between X and Y given Z is defined as

$$I(X;Y\,|\,Z) = \mathrm{E}\left[\ln\left(\frac{p(X,Y\,|\,Z)}{p(X\,|\,Z)p(Y\,|\,Z)}\right)\right],$$

  and equivalently,

$$I(X;Y\,|\,Z) = \sum_{x,y,z} p(x,y,z) \ln\left(\frac{p(x,y\,|\,z)}{p(x\,|\,z)p(y\,|\,z)}\right) = \sum_z p(z) \sum_{x,y} p(x,y\,|\,z) \ln\left(\frac{p(x,y\,|\,z)}{p(x\,|\,z)p(y\,|\,z)}\right).$$

- By the information inequality,

$$I(X;Y\,|\,Z) \ge 0, \quad \text{and} \quad I(X;Y\,|\,Z) = 0 \ \text{ iff. } \ X \perp Y \mid Z.$$

Implications of Information Inequality

for Maximum Likelihood Estimation

Setup for discussion:

- Let $y_1^n = (y_1, y_2, \ldots, y_n)$ be the observed values of a random sample of size n, $Y_1, Y_2, \ldots, Y_n$. Assume that each Y_i is distributed as Y_0 and that the true distribution is Q^* with the PMF q^*.
- Let Q_n be the empirical distribution of Y_0, given by the observed frequencies in the data y_1^n. Let q_n denote the PMF.
- For our model Θ, let p_θ denote the PMF of Y_0 corresponding to θ, and let $\hat{\theta}_n$ denote the maximum likelihood estimate based on y_1^n.

Let θ^* correspond to the distribution in Θ that is closest to Q^* in terms of KL-divergence (assume θ^* exists):

$$\theta^* \in \arg\min_{\theta\in\Theta} \mathrm{KL}\big(q^*, p_\theta\big) = \arg\min_{\theta\in\Theta} \mathrm{E}_{Q^*}\big[-\ln p(Y_0;\theta)\big].$$

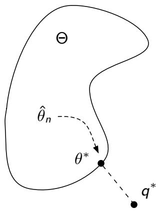
(The equality above follows from

$$\mathrm{KL}\big(q^*, p_\theta\big) = \mathrm{E}_{Q^*}\left[\ln\left(\frac{q^*(Y_0)}{p(Y_0;\theta)}\right)\right] = \mathrm{E}_{Q^*}\big[-\ln p(Y_0;\theta)\big] - H(Y_0)$$

and the fact that the entropy term $H(Y_0)$ is a constant independent of θ.)

Implications of Information Inequality

for Maximum Likelihood Estimation



Under mild conditions, as $n \to \infty$, $Q_n \to Q^*$, $\hat{\theta}_n \to \theta^*$, and

$$\begin{aligned} -\ell(\hat{\theta}_n; y_1^n) &= n\,\mathrm{E}_{Q_n}\big[-\ln p(Y_0;\hat{\theta}_n)\big] \\ &\approx n\,\mathrm{E}_{Q^*}\big[-\ln p(Y_0;\hat{\theta}_n)\big] + o(n) \\ &\approx n\,\mathrm{KL}\big(q^*, p_{\theta^*}\big) + nH(Y_0) + o(n). \end{aligned}$$

We can always distinguish between a fixed correct model Θ_1 and a fixed wrong model Θ_2 with enough data, because

$$\begin{aligned} &\exists\,\theta \in \Theta_1,\ s.t.\ \ p_\theta = q^*, \qquad \text{(def. of a correct model)} \\ &\qquad \Rightarrow\ \ \mathrm{KL}\big(q^*, p_{\theta_1^*}\big) = 0; \\ &\nexists\,\theta \in \Theta_2,\ s.t.\ \ p_\theta = q^*, \qquad \text{(def. of a \textit{mis-specified} -- wrong -- model)} \\ &\qquad \Rightarrow\ \ \mathrm{KL}\big(q^*, p_{\theta_2^*}\big) > 0. \end{aligned}$$

(We assume θ_2^* exists in the above.)

Outline

Occam's Razor – Principle of Parsimony

Occam's razor:

- William of Ockham or Occam (?1285-1937/1349) is regarded as one of the most important philosophers of his time.
- Occam's razor refers to the principle of parsimony: 'it is vain to do with more what can be done with fewer.'
- Apply the principle to model selection: We favor simple models over complex ones if they fit data about equally well. (But what does "about equally well" mean?)

Informal discussion:

- If models $\Theta_1, \Theta_2, \cdots$ are all correct, then $\min_{\theta \in \Theta_i} \mathrm{KL}(q^*, p_\theta) = 0$ for all Θ_i, so on this basis they are all indistinguishable from the true model. Following the parsimony principle, we would prefer the simplest model.
- An observation from a different viewpoint: We have

  $$\mathrm{KL}(q^*, p_{\theta^*}) = \min_{\theta \in \Theta} \mathrm{KL}(q^*, p_\theta) \le \mathrm{KL}(q^*, p_{\hat{\theta}}).$$

  Adding more parameters to the model decreases $\mathrm{KL}(q^*, p_{\theta^*})$. But with finite samples, such decrease may be outweighed by the increase in $\mathrm{KL}(q^*, p_{\hat{\theta}})$. This suggests that we may compare models based on

  $$\mathrm{E}\big[\mathrm{KL}(q^*, p_{\hat{\theta}})\big],$$

  where the expectation is with respect to the true distribution of the random sample that gives rise to $\hat{\theta}$.

Two Likelihood Criteria for Model Selection

Suppose that Θ has d free parameters. Let $\hat{\theta}$ be the maximum likelihood estimate of θ based on the observed values of a random sample of size n.

Akaike's information criterion (AIC) and Bayes information criterion (BIC):

$$\mathrm{AIC} = -2\,\ell(\hat{\theta}) + 2d, \qquad \mathrm{BIC} = -2\,\ell(\hat{\theta}) + 2d \ln n.$$

Model selection with AIC/BIC:
calculate the AIC/BIC scores for models $\Theta_1, \Theta_2, \ldots$; select the model with the minimal score (or consider several near-optimal models)

Notes:

- Both AIC and BIC can be viewed as crude approximations of the quantity $2n\,\mathrm{E}\big[\mathrm{KL}(q^*, p_{\hat{\theta}})\big] + 2n\,c$, where c is some constant.
- AIC is inconsistent in the sense that if both the true, simpler model and a correct model are fitted, the probability of selecting the true model does not approach 1 as $n \to \infty$.
- BIC is consistent in the above sense, but with finite samples it tends to suggest a too parsimunous model and leads to underfit.
- Both criteria are used in practice beyond random samples.
- Model selection criteria continue to be an important research topic.

Further Readings

For an overview of likelihood and model selection with likelihood criteria:

1. A. C. Davison. *Statistical Models*, Cambridge Univ. Press, 2003. Chap. 4.1, 4.7.

Announcement:
The first three books in the reference list given in the first lecture are now in the reading room of the Kumpula library (1st floor). Books there are ordered alphabetically by authors' names.