Infotheory for Statistics and Learning

Lecture 14

- I-projections [CT:10.8], [CTu]
- Convergence of iterative projections [CTu], [C1]
- Maximum likelihood as a projection [C2]
- The EM algorithm [C2]

I-projections

Assume $\mathcal{P}$ and $\mathcal{Q}$ are convex sets of probability measures on $(\Omega, \mathcal{A})$
i.e., for $\mathcal{P}$; P and P' in $\mathcal{P} \Rightarrow \gamma P + (1-\gamma)P' \in \mathcal{P}$ for any $\gamma \in (0,1)$

For any R on $(\Omega, \mathcal{A})$, if there is a $P^* \in \mathcal{P}$ such that

$$\inf_{P \in \mathcal{P}} D(P\|R) = D(P^*\|R)$$

then P^* is an I-projection of R on $\mathcal{P}$, notation $P^* = \Pi_{\mathcal{P}}(R)$

Similarly, if there is a $Q^* \in \mathcal{Q}$ such that

$$\inf_{Q \in \mathcal{Q}} D(R\|Q) = D(R\|Q^*)$$

then Q^* is a reverse I-projection of R on $\mathcal{Q}$, notation $Q^* = \bar{\Pi}_{\mathcal{Q}}(R)$

We also define

$$d(\mathcal{P}, \mathcal{Q}) = \inf_{P \in \mathcal{P}, Q \in \mathcal{Q}} D(P\|Q)$$

If $P^* = \Pi_{\mathcal{P}}(R)$ exists, then

$$D(P\|R) \geq D(P\|P^*) + D(P^*\|R)$$

for every $P \in \mathcal{P}$

If $Q^* = \bar{\Pi}_{\mathcal{Q}}(R)$ exists, then

$$D(P\|Q^*) \leq D(P\|R) + D(P\|Q)$$

for any P on $(\Omega, \mathcal{A})$ and every $Q \in \mathcal{Q}$

For an arbitrary Q_0 on $(\Omega, \mathcal{A})$, and with $P_1 = \Pi_{\mathcal{P}}(Q_0)$ and $Q_1 = \bar{\Pi}_{\mathcal{Q}}(P_1)$, we get

$$D(P\|Q) + D(P\|Q_0) \geq D(P\|Q_1) + D(P_1\|Q_1)$$

for every $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$

Proof of the first inequality: Let

$$p(\omega) = \frac{dP^*}{dR}, \quad q(\omega) = \frac{dP}{dR}$$

Since $P_t = (1-t)P + tP^* \in \mathcal{P}$ for each $t \in (0, 1]$

$$f(t) = D(P_t\|R)$$

is minimized at $t = 1$. Thus

$$0 \geq \frac{f(1) - f(t)}{1-t} = \int \frac{1}{1-t}(p \log p - p_t \log p_t) dR$$

where $p_t = (1-t)q + tp$. Letting $t \to 1$ we get

$$0 \geq \int (1 + \log p)(p - 1) dR = D(P^*\|R) - D(P\|R) + D(P\|P^*)$$

The proof of the second is similar, and the third follows from the first two

Lemma:

For real-valued sequences $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ and a number $c < \infty$ such that

$$c + b_{n-1} \geq b_n + a_n, \quad n = 1, 2, 3, \ldots$$

it holds that

$$\liminf_{n\to\infty} a_n \leq c$$

If in addition

$$\sum_{n=0}^{\infty} \max(c - a_n, 0) < \infty$$

we have

$$\lim_{n\to\infty} a_n = c$$

Iterative I- and reverse I-projections

Assume that $D(P\|Q) < \infty$ for all $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$, and that $d(\mathcal{P}, \mathcal{Q}) = D(P^*\|Q^*)$ ...can be generalized, see [CTu]

For an arbitrary $P_0 \in \mathcal{P}$, let

$$Q_0 = \bar{\Pi}_{\mathcal{Q}}(P_0),\ P_1 = \Pi_{\mathcal{P}}(Q_0),\ Q_1 = \bar{\Pi}_{\mathcal{Q}}(P_1), \ldots$$

then

$$\lim_{n\to\infty} D(P_n\|Q_n) = d(\mathcal{P}, \mathcal{Q})$$

Proof: For $P_{n+1} = \Pi_{\mathcal{P}}(Q_n)$ and $Q_{n+1} = \bar{\Pi}_{\mathcal{Q}}(P_{n+1})$ we have

$$D(P\|Q) + D(P\|Q_n) \geq D(P\|Q_{n+1}) + D(P_{n+1}\|Q_{n+1})$$

for all $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$

Apply the lemma with $c = D(P^*\|Q^*)$, $b_{n-1} = D(P^*\|Q_n)$, $b_n = D(P^*\|Q_{n+1})$ and $a_n = D(P_{n+1}\|Q_{n+1})$

I-projections for a finite Ω

Consider $P^* = \Pi_{\mathcal{P}}(R)$ for R on $(\Omega, \mathcal{A})$ and a convex $\mathcal{P}$

If for every $P \in \mathcal{P}$ there is a $\gamma \in (0,1)$ and a $P' \in \mathcal{P}$ such that

$$P^* = \gamma P + (1-\gamma)P'$$

then

$$D(P\|R) = D(P\|P^*) + D(P^*\|R) < \infty$$

for every $P \in \mathcal{P}$

For a finite Ω, the above is always true when $\mathcal{P}$ is linear in the sense that $Q = \gamma P + (1-\gamma)P' \in \mathcal{P}$ for any P and P' in $\mathcal{P}$ and all γ such that Q is a probability measure

For a finite Ω and with $\mathcal{P}_i$, $i = 1, \dots, k$, linear for each i and assuming $\mathcal{P} = \cap_i \mathcal{P}_i \neq \emptyset$, let R be any measure on $(\Omega, \mathcal{A})$ such that there is a $P \in \mathcal{P}$ for which $P \ll R$. Define $P_1 = \Pi_{\mathcal{P}_1}(R)$ and $P_{n+1} = \Pi_{\mathcal{P}_i}(P_n)$ for $n = mk + i$, $m = 0, 1, 2, \dots$, and $1 \leq i \leq k$

Let $P^* = \Pi_{\mathcal{P}}(R)$, then $\lim_{n\to\infty} D(P^*\|P_n) = 0$ and hence also

$$\lim_{n\to\infty} \mathsf{TV}(P^*, P_n) = 0$$

by Pinsker's inequality

Proof: We have $D(P^*\|P_{n-1}) = D(P^*\|P_n) + D(P_n\|P_{n-1})$ which gives

$$D(P^*\|R) = D(P^*\|P_n) + \sum_{i=1}^{n} D(P_i\|P_{i-1})$$

$\Rightarrow \lim_{n\to\infty} D(P_n\|P_{n-1}) = 0$. This then implies the result.

Maximum Likelihood

Assume $X_i \sim$ iid P for $i = 1, \dots, n$, and $X_i \in \mathcal{X}$ with $|\mathcal{X}| < \infty$

Take $\mathcal{X} = \{1, \dots, M\}$ for simplicity

For $X^n = (X_1, \dots, X_n)$ let $T_{x^n}(i)$ denote the type of $X^n = x^n$

For the pmf $p(i) = \Pr(X = i)$ of P, note that

$$\begin{aligned}\Pr(X^n = x^n) &= p(1)^{nT_{x^n}(1)} p(2)^{nT_{x^n}(2)} \cdots p(M)^{nT_{x^n}(M)} \\ &= \exp\left(n \sum_{i=1}^{M} T_{x^n}(i) \ln p(i)\right) \\ &= \exp\big[-n(H(T_{x^n}) + D(T_{x^n} \| p))\big]\end{aligned}$$

(with $H(T_{x^n})$ and $D(T_{x^n} \| p)$ in nats)

Assume p is unknown but it is known that $p \in \mathcal{P}$ for a convex and closed $\mathcal{P} \subset \mathbb{R}^M$ (for example the set of all pmf's)

Then, given $X^n = x^n$ the ML estimate of p is

$$p^* = \bar{\Pi}_{\mathcal{P}}(T_{x^n})$$

Expectation–Maximization (EM)

Consider $X \in \{1, 2, \ldots, K\}$ and $Y \in \{1, 2, \ldots, M\}$ jointly distributed according to P with pmf $p(i, j) = \Pr(X = i, Y = j)$

Assume we generate X^n and Y^n jointly iid $\sim P$ but only observe $Y^n = y^n$

- X is a latent or "hidden" variable

We wish to estimate P from Y^n

Assume it is known that $p \in \mathcal{P} \subset \mathbb{R}^{K \times M}$ for $\mathcal{P}$ convex and closed

Let $T_{X^n,y^n}(i, j)$ be the joint type for random X^n and observed y^n

Pick an arbitrary $q_0 \in \mathcal{P}$, let $\ell = 1$

Expectation (E) step: Set

$$T_\ell = E[T_{X^n,y^n} | Y^n = y^n]$$

assuming $q_{\ell-1}$ is the correct p

Maximization (M) step: Set q_ℓ equal to the ML estimate of P assuming T_ℓ is the joint type, T_{x^n,y^n}, for the full observation, i.e.

$$q_\ell = \bar{\Pi}_{\mathcal{P}}(T_\ell)$$

Repeat for $\ell = 2, 3, \ldots$

Note that

$$T_{x^n,y^n}(i,j) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}(\{x_k = i\})\mathbb{1}(\{y_k = j\})$$

Hence

$$\begin{aligned} E[T_{X^n,y^n}(i,j)|Y^n = y^n] &= \frac{1}{n}\sum_{k=1}^{n} \Pr(X_k = i|Y^n = y^n)\mathbb{1}(\{y_k = j\}) \\ &= \frac{p(i,j)}{p(j)} T_{y^n}(j) \end{aligned}$$

where $p(j) = \sum_i p(i,j)$ and $T_{y^n}(j) = n^{-1}\sum_k \mathbb{1}(\{y_k = j\})$

That is, for the E-step

$$T_\ell(i,j) = \frac{q_{\ell-1}(i,j)}{q_{\ell-1}(j)} T_{y^n}(j)$$

Since $T_\ell(i,j)/T_{y^n}(j) = T_\ell(i|j) = q_{\ell-1}(i,j)/q_{\ell-1}(j) = q_{\ell-1}(i|j)$ we get

$$\begin{aligned} D(T_\ell \| q_{\ell-1}) &= \sum_j T_{y^n}(j) \sum_i T_\ell(i|j) \ln \frac{T_\ell(i|j)}{q_{\ell-1}(i|j)} + D(T_{y^n}(j) \| q_{\ell-1}(j)) \\ &= 0 + D(T_{y^n}(j) \| q_{\ell-1}(j)) = \min_{T \in \mathcal{T}} D(T \| q_{\ell-1}) \end{aligned}$$

where $q_{\ell-1}(j) = \sum_i q_{\ell-1}(i,j)$ and

$$\mathcal{T} = \{\text{ types } T : \sum_i T(i,j) = T_{y^n}(j)\}$$

i.e. $T_\ell = \Pi_{\mathcal{T}}(q_{\ell-1})$

Consequently we have,

$$E\text{-step: } T_\ell = \Pi_{\mathcal{T}}(q_{\ell-1})$$

$$M\text{-step: } q_\ell = \bar{\Pi}_{\mathcal{P}}(T_\ell)$$