Spring 2017
STAT 364
Information Theory
Due: Feb 2, 2017 in class
Prof. Yihong Wu

1 Exercises

1. Bob is to eat all the cookies from a jar containing three peanut butter, two chocolate, and one oatmeal cookies. He decides to proceed completely randomly. Denote by X and Y the flavor of the first and the second cookie he eats.

   (a) Find $H(X)$, $H(Y)$, $H(X,Y)$, $H(Y|X)$, $H(X|Y)$, $I(X;Y)$, $D(P_{Y|X=\text{chocolate}} \| P_{Y|X=\text{oatmeal}})$ and $D(P_{Y|X=\text{oatmeal}} \| P_{Y|X=\text{chocolate}})$.

   (b) Now, what if Y denotes the flavor of the last cookie Bob eats?

   (c) How much information is contained in the sequence in which the cookies are eaten?

2. Let X be distributed according to the exponential distribution with mean $\mu > 0$, i.e., with density $p(x) = \frac{1}{\mu} e^{-x/\mu} \mathbf{1}_{\{x \geq 0\}}$. Let $a \in \mathbb{R}$. Compute the divergence $D(P_{X+a} \| P_X)$.

3. Let $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$ be the Gaussian distribution on $\mathbb{R}^n$ with mean $\mathbf{m} \in \mathbb{R}^n$ (column vector) and covariance matrix $\mathbf{\Sigma}$.

   (a) Let $\mathbf{I}_n$ be the $n \times n$ identity matrix. Show that

   $$D(\ \mathcal{N}(\mathbf{m}, \mathbf{I}_n) \ \| \ \mathcal{N}(0, \mathbf{I}_n)) = \frac{1}{2} \|\mathbf{m}\|_2^2$$

   (b) Show that

   $$D(\ \mathcal{N}(\mathbf{m}_1, \mathbf{I}_n) \ \| \ \mathcal{N}(\mathbf{m}_0, \mathbf{I}_n)) = \frac{1}{2} \|\mathbf{m}_0 - \mathbf{m}_1\|_2^2$$

   (Hint: think how Gaussian distribution changes under shifts $\mathbf{x} \mapsto \mathbf{x} + \mathbf{a}$. Apply data-processing to reduce to previous case.)

   (c) Assume that $\mathbf{\Sigma}$ is non-singular. Show that

   $$D(\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}) \| \mathcal{N}(\mathbf{m}_0, \mathbf{\Sigma}) = \frac{1}{2} (\mathbf{m}_0 - \mathbf{m}_1)^\top \mathbf{\Sigma}^{-1} (\mathbf{m}_0 - \mathbf{m}_1).$$

   (Hint: think how Gaussian distribution changes under non-singular linear transformations $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$. Apply data-processing to reduce to previous case.)

4. Recall that $d(p\|q) \triangleq D(\text{Bern}(p)\|\text{Bern}(q))$ denotes the binary divergence function:

   $$d(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}\,.$$

   (a) Prove for all $p, q \in [0,1]$

   $$d(p\|q) \geq 2(p-q)^2 \log e\,.$$

   Note: Proof by drawing is NOT accepted.

(b) Apply data processing inequality to prove the *Pinsker-Csiszár inequality*:

$$\mathrm{TV}(P,Q) \le \sqrt{\frac{1}{2\log e} D(P\|Q)}\,,$$

where $\mathrm{TV}(P,Q)$ is the *total variation* distance between probability distribution P and Q:

$$\mathrm{TV}(P,Q) \triangleq \sup_E \left(P[E] - Q[E]\right)\,,$$

with the supremum taken over all events E.

5. (a) Prove

$$2H(X,Y,Z) \le H(X,Y) + H(Y,Z) + H(Z,X)$$

(b) Use the above inequality to prove *Shearer's lemma*: Place n points in $\mathbb{R}^3$ arbitrarily. Let n_1, n_2, n_3 denote the number of distinct points projected onto the xy, xz and yz-plane, respectively. Then:

$$n_1 n_2 n_3 \ge n^2\,. \tag{1}$$

(c) What necessary conditions for equality in (1) can you state? Find explicit examples of equality (a single point does not count, please).

(d) Find examples where the left-hand side in (1) far exceeds the right-hand side.

Comments: This is an example of an information-theoretic proof of a combinatorial result.

2 Optional reading

1. Read [1, Chapter 1]
2. Read [2]
3. Watch `http://www.youtube.com/watch?v=z2Whj_nL-x8`

References

[1] T. Cover and J. Thomas, *Elements of Information Theory,* Second Edition, Wiley, 2006

[2] Solomon W. Golomb, Elwyn Berlekamp, Thomas M. Cover, Robert G. Gallager, James L. Massey, and Andrew J. Viterbi, *Claude Elwood Shannon (1916–2001)*, Notices of the AMS, Vol. 49, No. 1, 2002