Lecture 17

Mark Wilde

Sat Feb 4, 2017

§1

§2 Entropy of a Random Variable

Def^n: Info content (associated with a specific output) $:= i(x) = -\log(P_X(x))$

Def^n: Entropy : $H(X) := E_X\{i(X)\} = -\sum_x P_X(x) \log P_X(x)$

NB: Assume $\lim_{\varepsilon \to 0} \varepsilon \cdot \log(1/\varepsilon) = 0$ as justif^n for $0 \cdot \log(1/0) = 0$.

§2.1 Properties of Entropy

Prop 2 : (Non-negativity) : $H(X) \geq 0$ | Proof : $i(x) \geq 0$.

Prop 3 : (Concavity) : $H(X)$ is concave in the probability density $P_X(x)$.

Recall : A f^n $f : \mathbb{R}^n \to \mathbb{R}$ is convex $\Leftrightarrow$ dom f is convex & $\forall y \in \text{dom} f$, for $0 \leq \theta \leq 1$
$f(\theta x + (1-\theta) y) \leq \theta f(x) + (1-\theta) f(y)$

: We can check the derivatives (see Boyd page 72)
$f(x) = x \log x$ has $f'(x) = \log x + 1$, $f''(x) = 1/x > 0$ for $x > 0$
which establishes convexity. (which means $-f(x)$ is concave)

Prop 4 : (Permutat^n invariance) : <obvious>

Prop 5 : (Minimum value) : Entropy vanishes for deterministic variables

Prop 6 : (Max. value) : $H(X) \leq \log|X|$ where X is the alphabet set
: The saturation is achieved for a uniform random distr over X.

(Proof is given later)

§3 Conditional Entropy

Def^n : (Conditional Entropy) : $H(X|Y) := E_{XY}\{ i(X|Y) \}$
$= -\sum P_{X,Y}(x,y) \log(P_{X|Y}(x|y))$

Thm 8 : (Conditioning doesn't increase entropy)
: $H(X) \geq H(X|Y)$ (see rough page 2 for a proof)

Intuit^n : How much information about X Bob has if he knows Y.

§4 Joint Entropy

Def^n: (Joint Entropy): $H(X,Y) \equiv E_{xy}\{i(X,Y)\}$
$$= -\sum_{x,y} P_{X,Y}(x,y) \log(P_{X,Y}(x,y))$$

Intui^n: avg. Info gain upon learning both X & Y.

Mon, Feb 6, 2017

§5 Mutual Information

Def^n: (Mutual Inform^n): $I(X;Y) := H(X) - H(X|Y)$
$$= H(X) - H(X,Y) + H(Y)$$
$$= \sum_{x,y} P_{X,Y}(x,y) \log\left[\frac{P_{X,Y}(x,y)}{P_X(x)\,P_Y(y)}\right]$$

Intui^n: How much knowing Y reduces the uncertainty about X.

NB: Mutual Inf^n is symmetric $I(X;Y) = I(Y;X)$

Remark: Mutual Inf^n is zero when X & Y are independent.

Thm 12: $I(X;Y) \geq 0$

proof: Conditioning reduces H.

§6 Relative Entropy

Def^n: $\mathrm{supp}(f) := \{x : f(x) \neq 0\}$.

Def^n: (Relative Entropy) $D(p\|q) := \begin{cases} \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) & \text{if } \mathrm{supp}(p) \subseteq \mathrm{supp}(q) \\ +\infty & \text{else.} \end{cases}$

NB: $I(X;Y) = D\left(P_{X,Y}(x,y) \,\|\, P_X(x) P_Y(y)\right)$

Intui^n: How far is dist^n P from q

§7 Conditional Mutual Information

Def^n: (Conditional Mutual Information) $I(X;Y|Z) = H(Y|Z) - H(Y|X,Z)$
$$= H(X|Z) - H(X|Y,Z)$$
$$= H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

Thm 17 (Strong Subaditivity). $I(X;Y|Z) \geq 0$ & saturated iff $X - Z - Y$ is a Markov chain (i.e. $p_{X,Y|Z}(x,y|z) = p_{X|Z}(x)\, p_{Y|Z}(y)$)

proof: $I(X;Y|Z) = \sum_{z} p_Z(z)\, I(X;Y|Z=z) \geq 0 \;\because\; p_Z \geq 0$ & $I \geq 0$

satur^n from $I(X;Y) = 0$ iff $p(x,y) = p(x)\cdot p(y)$

Restatement: $H(XY|Z) \leq H(X|Z) + H(Y|Z)$

$H(XYZ) + H(Z) \leq H(XZ) + H(YZ)$

$H(X|YZ) \leq H(X|Z)$

Claim: $I(X_1, \ldots X_n; Y) = I(X_1;Y) + I(X_2;Y|X_1) + \ldots + (X_n;Y|X_1\ldots X_{n-1})$

§8 Entropy Inequalities

Thm 20 (Non-negativity of Relative Entropy). $p(x)$ is prob. over $\mathcal{X}$ alphabets $\sum_{q} q(x) \leq 1$ where $q : \mathcal{X} \to [0,1]$

$D(p\|q) \geq 0$ & $= 0$ iff $p = q$.

proof: For $\mathrm{supp}(p) \not\subseteq \mathrm{supp}(q)$ $D = +\infty$ so no issue

For $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$, use $\ln x \leq x - 1$ ($\forall\, x \geq 0$ & saturated at $x=1$)

$$D(p\|q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) = -\frac{1}{\ln 2} \sum_x p(x) \ln\left(\frac{q(x)}{p(x)}\right)$$

$$\geq +\frac{1}{\ln 2} \sum_x p(x)\left(1 - \frac{q(x)}{p(x)}\right) \qquad \text{using } -\ln x \geq 1 - x$$

$$= \frac{1}{\ln 2}\left(\sum p(x) - \sum q(x)\right)$$

$$\geq 0$$

To show further that equality entails $p = q$,

NB: $\sum p = \sum q = 1 \Rightarrow q$ is a prob.

Also $\ln\left(\frac{q}{p}\right) = 1 - \frac{q}{p}$ (saturated inequality!)

$\Rightarrow \frac{q}{p} = 1$ $\square$

Some proofs : For prop 6 (max value) : $0 \leq D\left(P_X(x) \,\|\, 1/|X|\right)$

$$= \sum_x P_X(x) \log\left(\frac{P_X(x)}{1/|X|}\right)$$

$$= -H(X) + \sum_x P_X(x) \log |X|$$

$$= -H(X) + \log|X|$$

$$\Rightarrow \quad H(X) \leq \log|X|.$$

For thm 8 : $I(X;Y) = D\left(P_{X,Y}(x,y) \,\|\, P_X(x)P_Y(y)\right) \geq 0$
$\overset{\shortparallel}{\text{thm } 12} \quad \Rightarrow \quad H(X) \geq H(X|Y)$

! Data processing inequalities & continuity of Entropy skipped for now !

Feb 7, 2017

§8.2 Data Processing Inequality

Remark : Two types (1) Correlations b/w two random variables only decrease if we process one of them (8.2.1)
(2) Relative entropy can't increase if a channel is applied to both the arguments. (8.2.2)

§8.2.1 Mutual Information Data-Processing Inequality

Intuition

Assume : Random variables X & Y, s.t. Y arises from X according to a stochastic map $\mathcal{N}_1 := P_{Y|X}(y|x)$

Recall : Mutual information captures correlations b/w random variables.

Assume : Random var. Z given by $\mathcal{N}_2 := P_{Z|Y}(z|y)$

Statement : $I(X;Y) \geq I(X;Z)$; Correlations b/w X & Z must be less than those b/w X & Y.

NB 1 : Stochastic maps subsume deterministic maps.

NB 2 : Z depends on Y only & not on X ; $X \rightarrow \boxed{\mathcal{N}_1} \rightarrow Y \rightarrow \boxed{\mathcal{N}_2} \rightarrow Z$

$$P_{Z|Y,X}(z|y,x) = P_{Z|Y}(z|y)$$

Def^n : Markov Chain $\overset{\text{impl.}}{:=}$ X, Y, Z form a markov chain $\Leftrightarrow X \rightarrow Y \rightarrow Z$.

Thm 21 (Data processing inequality). For $X \to Y \to Z$ (Markov Chain)

$$I(X;Y) \geq I(X;Z)$$

proof. $\because$ $X \to Y \to Z$, $P_{X,Z|Y}(x,z|y) = P_{Z|Y,X}(z|y,x)\, P_{X|Y}(x|y)$

$$= P_{Z|Y}(z|y)\, P_{X|Y}(x|y)$$

Consider: $I(X;YZ) = I(X;Y) + I(X;Z|Y)$ (using chain rule)

$= I(X;Y)$ (from thm 17, $I(X;Z|Y) = 0$ if $X \to Y \to Z$).

Also $I(X;YZ) = I(X;Z) + I(X;Y|Z)$

$\Rightarrow$ $I(X;Y) = I(X;Z) + I(X;Y|Z)$

Thm follows since (from thm 17) $I(X;Y|Z) \geq 0$.

$\square$

Corollary 22. $I(X;Y) \geq I(X;Y|Z)$

proof. almost as above. $\square$

§ 8.2.2 Relative Entropy Data-Processing Inequalities

Remark: The inequality here, follows from non-negativity of relative entropy.

Corollary 23 (Monotonicity of Relative Entropy).

Let p be a prob. on alphabet $\mathcal{X}$,

$q : \mathcal{X} \to [0,\infty)$

Let $N(y|x)$ be a conditional prob. distr.

Then $$D(p\|q) \geq D(Np\|Nq)$$

where $Np(y) = \sum_x N(y|x)\, p(x)$

$Nq(y) = \sum_x N(y|x)\, q(x)$.

Not clear

Saturation viz. $D(p\|q) = D(Np\|Nq)$ is achieved for a channel R defined as

$$R(x|y)\,(Nq)(y) = N(y|x)\, q(x)$$

& R is s.t. $RNp = p$, viz. $RNp(x) = \sum_{y,x'} R(x|y)\, N(y|x')\, p(x')$

proof. For p, q s.t. $supp(p) \not\subseteq supp(q)$, $D = +\infty$ & (inequality is satisfied.

For p, q s.t. $supp(p) \subseteq supp(q)$, $\Rightarrow supp(Np) \subseteq supp(Nq)$

we have

$$
\begin{aligned}
& D(Np \| Nq) \\
&= \sum_y (Np)(y) \log \left[\frac{(Np)(y)}{(Nq)(y)}\right] \\
&= \sum_{y,x} N(y|x)\, p(x) \log \left[\frac{(Np)(y)}{(Nq)(y)}\right] \\
&= \sum_x p(x) \left[\sum_y N(y|x) \log \left[\frac{(Np)(y)}{(Nq)(y)}\right]\right] \\
&= \sum_x p(x) \ln e^{\sum_y N(y|x) \log \frac{Np(y)}{Nq(y)}}
\end{aligned}
$$

$$\Rightarrow \quad D(p\|q) - D(Np\|Nq) = D(p\|r)$$

for $r := q(x)\, e^{\left(\sum_y N(y|x) \log \left(\frac{Np(y)}{Nq(y)}\right)\right)}$

$$
\begin{aligned}
\Rightarrow \sum_x r(x) &\leq \sum_x q(x) \sum_y N(y|x) \exp\log\left[\frac{Np(y)}{Nq(y)}\right] \quad \text{(using convexity of the exponential } f^n) \\
&= \sum_x q(x) \sum_y N(y|x) \frac{Np(y)}{Nq(y)} \\
&= \sum_y \left[\cancel{\sum_x q(x) N(y|x)}\right] \frac{Np(y)}{\cancel{Nq(y)}} \\
&= \sum_y Np(y) \\
&= 1
\end{aligned}
$$

Since $\sum_x r(x) \leq 1$, from thm 20, $D(p\|r) \geq 0$

which proves the thm, $D(p\|q) \geq D(Np\|Nq)$.

Not clear, Nor complete

About saturation: Assume $RNp = p$

$$D(Np\|Nq) \geq D(RNp\|RNq) = D(p\|q)$$

using the aforesaid

Rough

$\log_A C = d$

$\cancel{A} \; C = A^d$

$\ln C = d \ln A$

$\frac{\ln C}{\ln A} = \log_A C$

$p : X \to [0,1] \quad \Sigma p = 1$

$q : X \to [0, \infty)$

$N(y|x)$

$D(p\|q) \geq D(Np\|Nq)$

$Np = (Np)(y) := \sum_x N(y|x)\, p(x)$

$Nq = Nq(y) := \sum_x N(y|x)\, q(x)$

$R(x|y)(Nq)(y) = N(y|x)\, q(x)$

$D(Np\|Nq)$

$= \sum_y (Np)(y) \log \left[\frac{(Np)(y)}{(Nq)(y)} \right]$

$N(y|x)\, p(x) \log$

$=$

$p(x) \log \frac{p(x)}{q(x)}$

$\log \frac{p(x)}{(q(x)\, e \ldots\ )}$

$R =$