Optimization

Mahdi Roozbahani
Georgia Tech

Outline

Motivation

Entropy

Conditional Entropy and Mutual Information

Cross-Entropy and KL-Divergence

Let’s work on this subject in our Optimization lecture

Cross Entropy

Cross Entropy: The expected number of bits when a wrong distribution Q is assumed while the data actually follows a distribution P

$$H(p,q) = -\sum_{x\in\mathcal{X}} \overbrace{p(x)}^{\text{actual pdf}} \, \log \overbrace{q(x)}^{\text{predicted pdf}} = H(P) + KL[P][Q]$$

This is because:

$$H(p,q) = \mathrm{E}_p[l_i] = \mathrm{E}_p\left[\log \frac{1}{q(x_i)}\right]$$

$$H(p,q) = \sum_{x_i} p(x_i)\, \log \frac{1}{q(x_i)}$$

$$H(p,q) = -\sum_{x} p(x)\, \log q(x).$$

Labeling target values
Label encoding (ordinal) and One-hot encoding

$$X = \begin{array}{c} h \quad w \quad a \quad \cdots \\ \left[\quad \quad \quad \quad \right]_{n \times d} \end{array}$$

$$\underset{\text{actual}}{Y} = \begin{bmatrix} cat \\ dog \\ fish \\ cat \\ \vdots \end{bmatrix}_{n \times 1} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \\ \vdots \end{bmatrix} \qquad \hat{Y}_{\text{Predicted}} = \begin{bmatrix} 1 \\ 2 \\ 2.5 \\ 1.5 \\ \vdots \end{bmatrix}$$

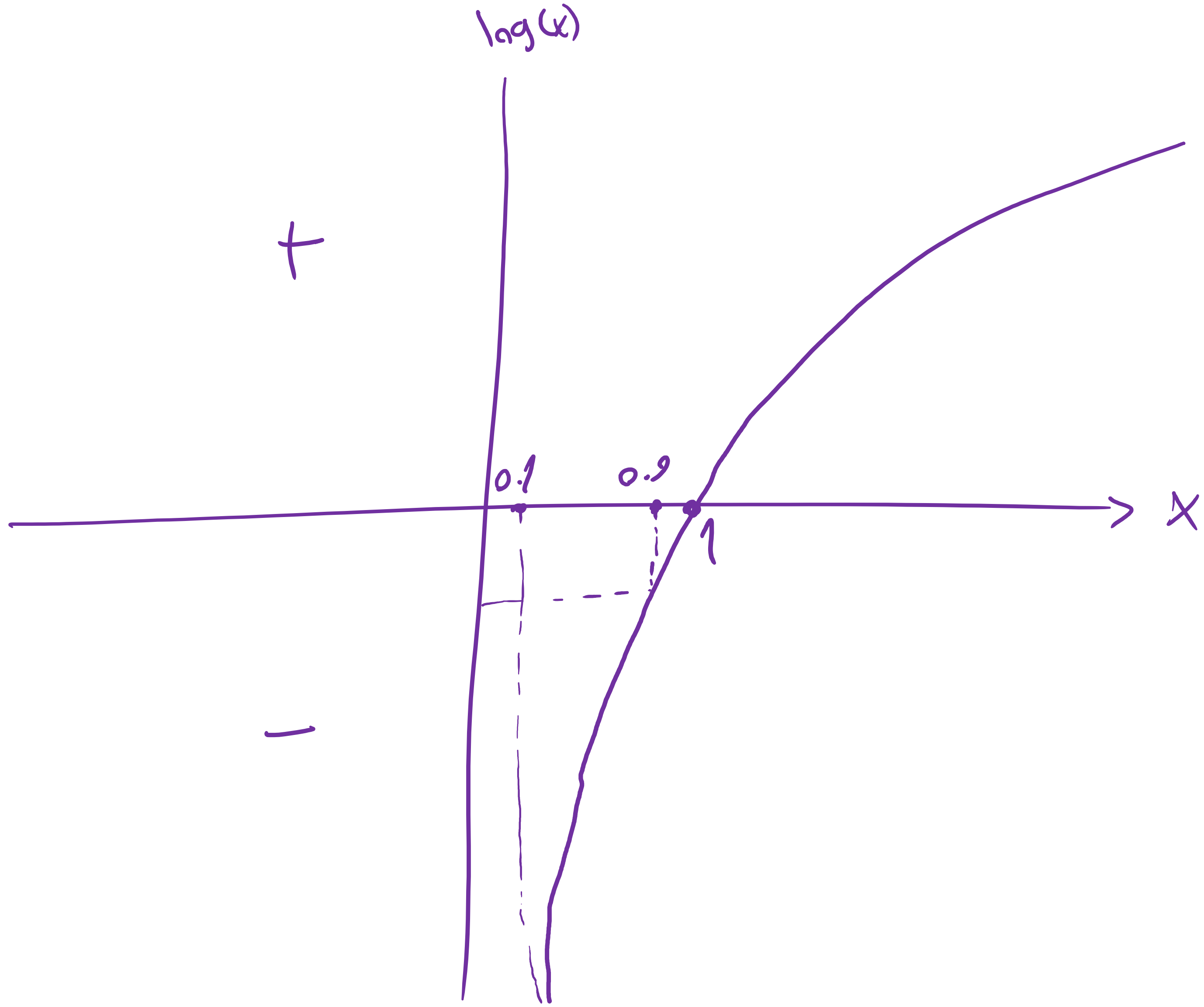
$$cat = [1 \quad 0 \quad 0]$$
$$dog = [0 \quad 1 \quad 0]$$
$$fish = [0 \quad 0 \quad 1]$$

$$Y_{\text{actual}} = \begin{bmatrix} [1 & 0 & 0] \\ [0 & 1 & 0] \\ [0 & 0 & 1] \\ [1 & 0 & 0] \\ & \vdots & \end{bmatrix} \qquad \hat{Y}_{\text{Predicted}} = \begin{bmatrix} [0.8 & 0.1 & 0.1] \\ [0.3 & 0.6 & 0.1] \\ [0.1 & 0.2 & 0.7] \\ & \vdots & \end{bmatrix}$$

$$\text{Loss (dot product)} = [1 \quad 0 \quad 0][0.8 \quad 0.1 \quad 0.1] + [0 \quad 1 \quad 0][0.3 \quad 0.6 \quad 0.1] + \cdots -$$

$$CE = H(p, q) = -\sum p(x) \log_2 q(x) = -[1 \log_2 0.8 + 0 \log_2 0.1 + 0 \log_2 0.1] - [0 \log_2 0.3 + 1 \log_2 0.6 + \cdots]$$

Why Cross entropy and not simply use dot product?



log 1 = 0

log 0.9

log 0.1

Kullback-Leibler Divergence

Another useful information theoretic quantity measures the difference between two distributions.

$$\mathbf{KL}[P(S)\|Q(S)] = \sum_s P(s) \log \frac{P(s)}{Q(s)}$$

$$= \underbrace{\sum_s P(s) \log \frac{1}{Q(s)}}_{\text{cross entropy}} - \mathbf{H}[P] = H(P,Q) - H(P)$$

Excess cost in bits paid by encoding according to Q instead of P.

KL Divergence is a **KIND OF** distance measurement

$$-\mathbf{KL}[P\|Q] = \sum_s P(s) \log \frac{Q(s)}{P(s)}$$

log function is concave or convex?

$$\sum_s P(s) \log \frac{Q(s)}{P(s)} \leq \log \sum_s P(s) \frac{Q(s)}{P(s)}$$

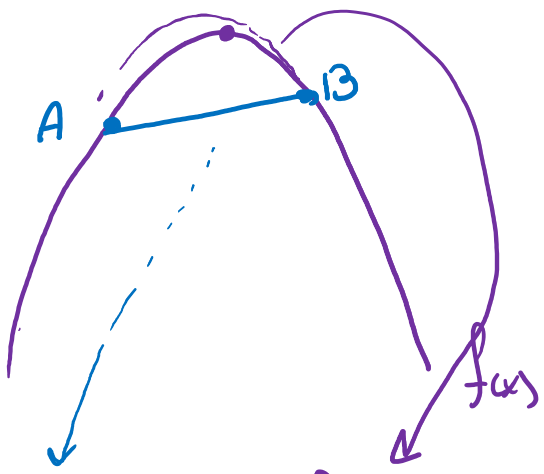
$$= \log \sum_s Q(s) = \log 1 = 0$$

By Jensen Inequality

So $\mathbf{KL}[P\|Q] \geq 0$. Equality iff $P = Q$

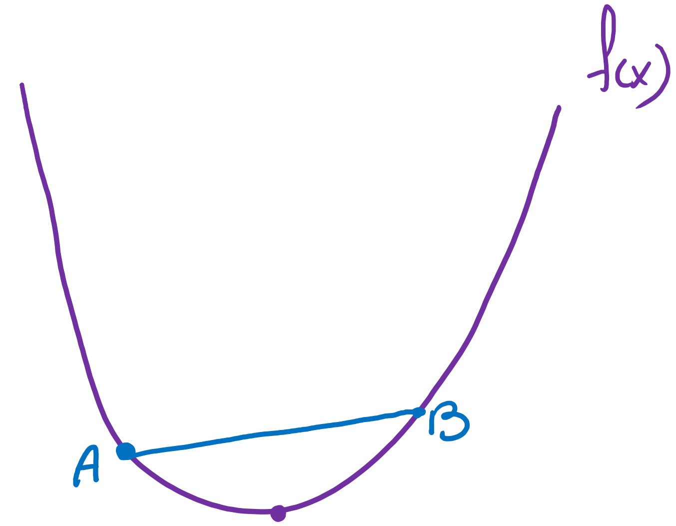
When $P = Q,\ KL[P||Q] = 0$

Concave



$$E[f(x)] < f(E[x])$$

Convex

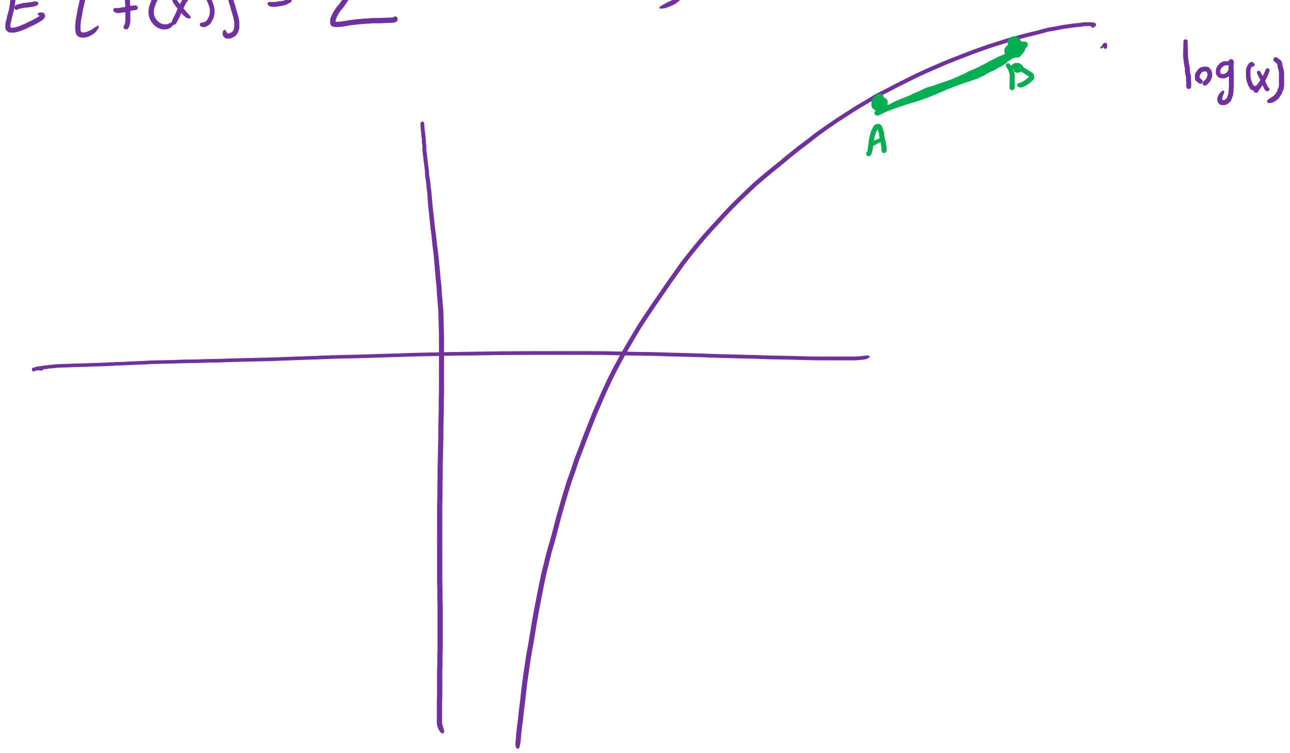


$$f(x) = x^2$$

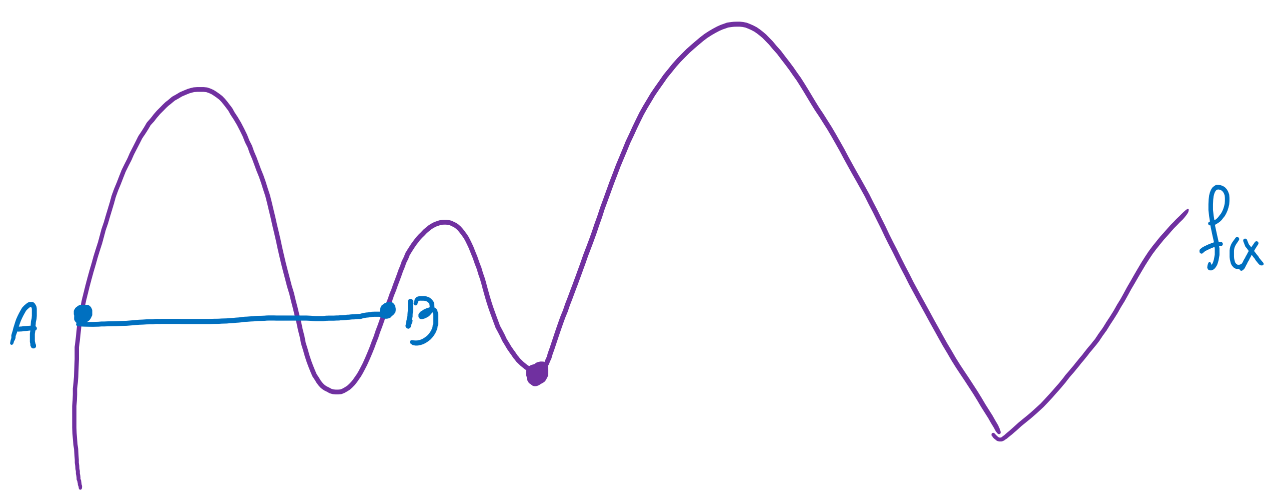
$$f''(x) = 2$$

$$E[f(x)] > f(E[x])$$

$$E[f(x)] = \sum p(x) f(x)$$



$$E[\log(x)] < \log(E[x])$$



$$-KL[P][Q] = \sum P(x) \log \underbrace{\frac{Q(x)}{P(x)}}_{g(x)} = \sum P(x) \log g(x) = E[\log g(x)]$$

$$-KL[P][Q] = E[\log g(x)] \leq \log(E[g(x)])$$

$$\text{- - - -} \leq \log\left(\sum P(x)\, g(x)\right)$$

$$\text{- - -} \leq \log\left(\sum \cancel{P(x)} \frac{Q(x)}{\cancel{P(x)}}\right)$$

$$\text{- - -} \leq \log\left(\sum Q(x)\right)$$

$$-KL[P][Q] \leq \log(1) = 0$$

$$KL[P][Q] \geq 0$$

Optimization

1. Probabilistic models $\rightsquigarrow$ Gaussian distribution $\begin{cases} \mu \\ \sigma \end{cases} \Rightarrow \underbrace{\text{MLE}}$

$\downarrow$

take a derivative

2. Non-probabilistic models $\rightsquigarrow$ constraints $\begin{cases} \text{Equality} \rightsquigarrow \text{Lagrange multiplier} \\ \text{Inequality} \rightsquigarrow \text{KKT} \end{cases}$

$$f(M,S) = 6M^2 + 3S^2$$

M : # hours you study ML per day
S : # " you sleep " "

$$\frac{\partial f(M,S)}{\partial M} = 0 \Rightarrow 12M = 0 \Rightarrow M = 0$$

$$\frac{\partial f(M,S)}{\partial S} = 0 \Rightarrow 6S = 0 \Rightarrow S = 0$$

$f(M,S) = 6M^2 + 3S^2$ ⇝ objective function

s.t. $M+S = 24$ ⇝ $g(M,S) = M+S-24$

$$L(M,S,\lambda) = f(M,S) - \lambda\, g(M,S)$$

$f(M,S)$ $\lambda, \alpha, \beta, \ldots$

$g(M,S) = M+S-24$

$h(M,S) = M-S-8$

what is you have two constraints

$L = (M,S,\lambda_1,\lambda_2) = f(M,S) - \lambda_1 g(M,S) - \lambda_2 h(M,S)$

$\nabla f(M,S) = \nabla g(M,S)$

$\nabla f(M,S) - \nabla g(M,S) = 0$

$$L(M,S,\lambda) = 6M^2 + 3S^2 - \lambda(M+S-24)$$

$$\frac{\partial L(M,S,\lambda)}{\partial \lambda} = 0 \Rightarrow M+S = 24 \Rightarrow \frac{\lambda}{12} + \frac{\lambda}{6} = 24 \Rightarrow \lambda = 96$$

$$\frac{\partial L(M,S,\lambda)}{\partial M} = 0 \Rightarrow 12M - \lambda = 0 \Rightarrow M = \frac{\lambda}{12} = \frac{96}{12} = 8$$

$$\frac{\partial L(M,S,\lambda)}{\partial S} = 0 \Rightarrow 6S - \lambda = 0 \Rightarrow S = \frac{\lambda}{6} = \frac{96}{6} = 16$$

$$M+S = 8+16 = 24 \checkmark$$

$f(M,S) = 6M^2 + 3S^2$

s.t. $M + S \leq 24 \Rightarrow g(M,S) = M + S - 24$

$g(M,S) < 0$

$L(M,S,\delta) = f(M,S) - \delta\, g(M,S)$

We need to satisfy $\underline{\underline{4}}$ conditions

(1) Stationary condition

(2) Primal feasibility $g(M,S) < 0$

(3) Dual feasibility $\delta \geq 0$

(4) Complementary slackness $g(M,S)\,\delta = 0$

- $\delta = 0$ then $g(M,S) \neq 0$
- $\delta \neq 0$ then $g(M,S) = 0$

Example 1:

https://www.geogebra.org/3d/srzmv8uh

Example 2:

https://www.geogebra.org/3d/syhkqpk7