CS 412 Introduction to Machine Learning

Probability Theory (2)

Instructor: Wei Tang

Department of Computer Science
University of Illinois at Chicago
Chicago IL 60607

https://tangw.people.uic.edu
tangw@uic.edu

Slides credit: Sargur N. Srihari

Announcement

- Machine problem #1 due on 9/29 (next Wes)
- Carefully follow the instructions

Topics

1. Entropy as an Information Measure
    1. Discrete variable definition
        Relationship to Code Length
    2. Continuous Variable
        Differential Entropy
2. Maximum Entropy
3. Conditional Entropy
4. Kullback-Leibler Divergence (Relative Entropy)
5. Mutual Information

Information Measure

- How much information is received when we observe a specific value for a discrete random variable x?
- Amount of information is degree of surprise
  - Certain means no information
  - More information when event is unlikely
- Depends on probability distribution $p(x)$, a quantity $h(x)$
- If there are two unrelated events x and y we want $h(x,y)= h(x) + h(y)$
- Thus we choose $h(x)= - \log_2 p(x)$
  - Negative assures that information measure is positive
- Average amount of information transmitted is the expectation wrt $p(x)$ refered to as entropy

$$H(x)=-\sum_x p(x) \log_2 p(x)$$

Usefulness of Entropy

- Uniform Distribution
  - Random variable x has 8 possible states, each equally likely
    - We would need 3 bits to transmit
    - Also, $H(x) = -8 \times (1/8)\log_2(1/8) = 3$ bits
- Non-uniform Distribution
  - If x has 8 states with probabilities

    $(1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64)$

    $H(x) = 2$ bits

- Non-uniform distribution has smaller entropy than uniform
- Has an interpretation of in terms of disorder

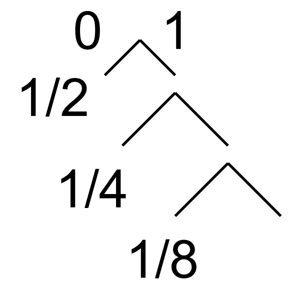
Relationship of Entropy to Code Length

- Take advantage of non-uniform distribution to use shorter codes for more probable events
- If x has 8 states (a,b,c,d,e,f,g,h) with probabilities

  $(1/2,1/4,1/8,1/16,1/64,1/64,1/64,1/64)$

  Can use codes $0,10,110,1110,111100,\ 111101,111110,111111$

  $$\text{average code length} = (1/2)1+(1/4)2+(1/8)3+(1/16)4+4(1/64)6 = 2 \text{ bits}$$



- Same as entropy of the random variable
- Shorter code string is not possible due to need to disambiguate string into component parts
- 11001110 is uniquely decoded as sequence *cad*

https://en.wikipedia.org/wiki/Huffman_coding

Relationship between Entropy and Shortest Coding Length

- Noiseless coding theorem of Shannon
  - Entropy is a lower bound on number of bits needed to transmit a random variable

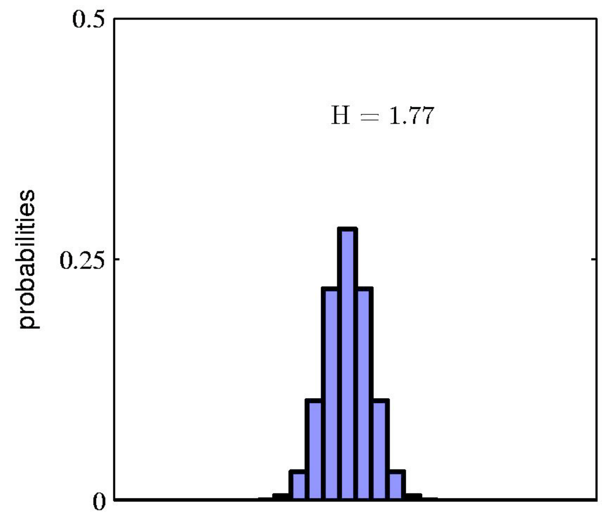
Entropy and Histograms

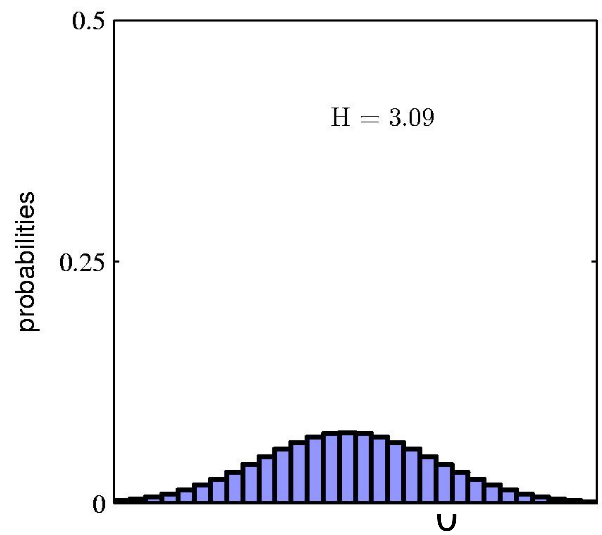
- If X can take one of M values (bins, states) and $p(X=x_i)=p_i$ then

$H(p)=-\sum_i p_i \ln p_i$

- Minimum value of entropy is 0 when one of the $p_i=1$ and other p_i are 0
  - noting that $\lim_{p\to 0} p \ln p = 0$
- Sharply peaked distribution has low entropy
- Distribution spread more evenly will have higher entropy

30 bins, higher value for broader distribution





Entropy with Multiple Continuous Variables

- Differential Entropy for multiple continuous variables

$$H(\mathrm{x}) = -\int p(\mathrm{x}) \ln p(\mathrm{x}) d\mathrm{x}$$

- For what distribution is differential entropy maximized?
  - For discrete distribution, it is uniform
  - For continuous, it is Gaussian

Principle of maximum entropy

- The probability distribution which best represents the current state of knowledge about a system is the one with largest entropy.

- Consider the set of all trial probability distributions that would encode the prior data.

- According to this principle, the distribution with maximal information entropy is the best choice.

- Another reason why most commonly we choose Gaussian!

Conditional Entropy

- If we have joint distribution $p(\mathrm{x},\mathrm{y})$
  - We draw pairs of values of x and y
  - If value of x is already known, additional information to specify corresponding value of y is $-\ln p(\mathrm{y}|\mathrm{x})$
- Average additional information needed to specify y is the conditional entropy

$$H[\mathrm{y}\,|\,\mathrm{x}] = -\iint p(\mathrm{y},\mathrm{x}) \ln p(\mathrm{y}\,|\,\mathrm{x})\, d\mathrm{y}\, d\mathrm{x}$$

- By product rule $\boxed{\mathrm{H}[\mathrm{x},\mathrm{y}] = \mathrm{H}[\mathrm{y}|\mathrm{x}] + \mathrm{H}[\mathrm{x}]}$
  - where H[x,y] is entropy of $p(\mathrm{x},\mathrm{y})$
  - H[x] is entropy of p(x)
  - Information needed to describe x and y is given by

    information needed to describe x plus

    additional information needed to specify y given x

Cross Entropy

- If we have modeled unknown distribution $p(\mathrm{x})$ by approximating distribution $q(\mathrm{x})$
  - i.e., $q(\mathrm{x})$ is used to construct a coding scheme of transmitting values of x to a receiver
  - Average amount of information required to specify value of x as a result of using $\mathrm{q(x)}$ instead of true distribution $\mathrm{p(x)}$ is given by cross entropy

$$H(p,q) = -\int p(x)\ln q(x)\,dx$$

Relative Entropy

- If we have modeled unknown distribution $p(\mathrm{x})$ by approximating distribution $q(\mathrm{x})$
  - i.e., $q(\mathrm{x})$ is used to construct a coding scheme of transmitting values of x to a receiver
  - Average additional amount of information required to specify value of x as a result of using $\mathrm{q}(\mathrm{x})$ instead of true distribution $\mathrm{p}(\mathrm{x})$ is given by relative entropy or K-L divergence
- Important concept in Bayesian analysis
  - Entropy comes from Information Theory
  - *K-L Divergence*, or *relative entropy*, comes from Pattern Recognition, since it is a distance (dissimilarity) measure

Relative Entropy or K-L Divergence

- Additional information required as a result of using $q(\mathrm{x})$ in place of $p(\mathrm{x})$

$$KL(p\,||\,q) = -\int p(x)\ln q(x)\,dx - \left(-\int p(x)\ln p(x)\,dx\right)$$

$$H(p,q) = H(p) + D_{\mathrm{KL}}(p\|q)$$

- Not a symmetrical quantity: $KL(p||q) \neq KL(q||p)$
- K-L divergence satisfies $KL(p||q)>0$ with equality iff $p(x)=q(x)$

Cross Entropy versus KL Divergence

$$H(p, q) = H(p) + D_{\mathrm{KL}}(p \| q)$$

Mutual Information

- Given joint distribution of two sets of variables $p(\mathrm{x},\mathrm{y})$
  - If independent, will factorize as $p(\mathrm{x},\mathrm{y}) = p(\mathrm{x})p(\mathrm{y})$
  - If not independent, whether close to independent is given by
    - KL divergence between joint and product of marginals

$$I[\mathrm{x},\mathrm{y}] = KL(p(\mathrm{x},\mathrm{y}) \,\|\, p(\mathrm{x})p(\mathrm{y}))$$

$$= \iint p(\mathrm{x},\mathrm{y}) \ln\left(\frac{p(\mathrm{x})p(\mathrm{y})}{p(\mathrm{x},\mathrm{y})}\right) d\mathrm{x}\, d\mathrm{y}$$

    - Called Mutual Information between variables x and y

Mutual Information

- Using Sum and Product Rules

  $$I[\mathrm{x},\mathrm{y}] = H[\mathrm{x}] - H[\mathrm{x}|\mathrm{y}] = H[\mathrm{y}] - H[\mathrm{y}|\mathrm{x}]$$

  - Mutual Information is reduction in uncertainty about x given value of y (or vice versa)

- Bayesian perspective:
  - if $p(\mathrm{x})$ is prior and $p(\mathrm{x}|\mathrm{y})$ is posterior, mutual information is reduction in uncertainty after y is observed