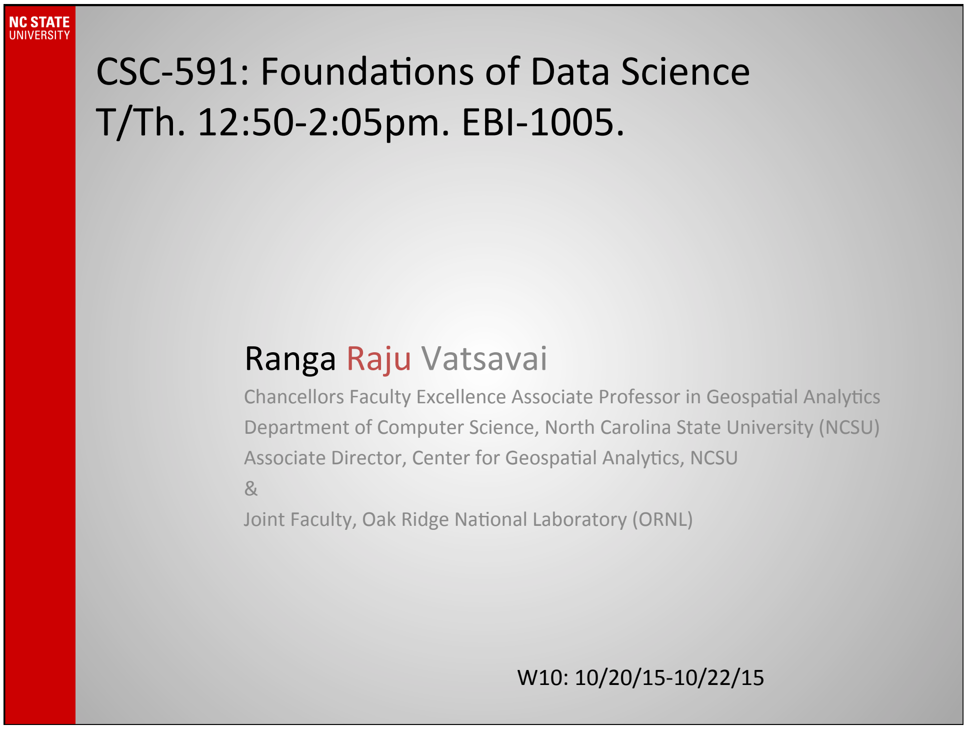
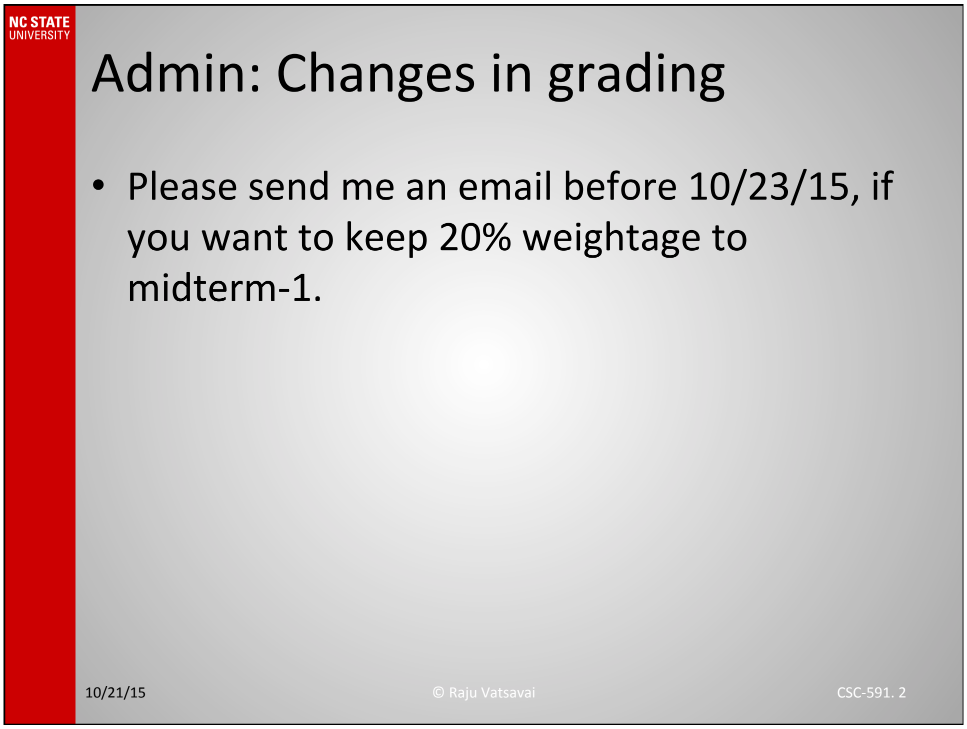


NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

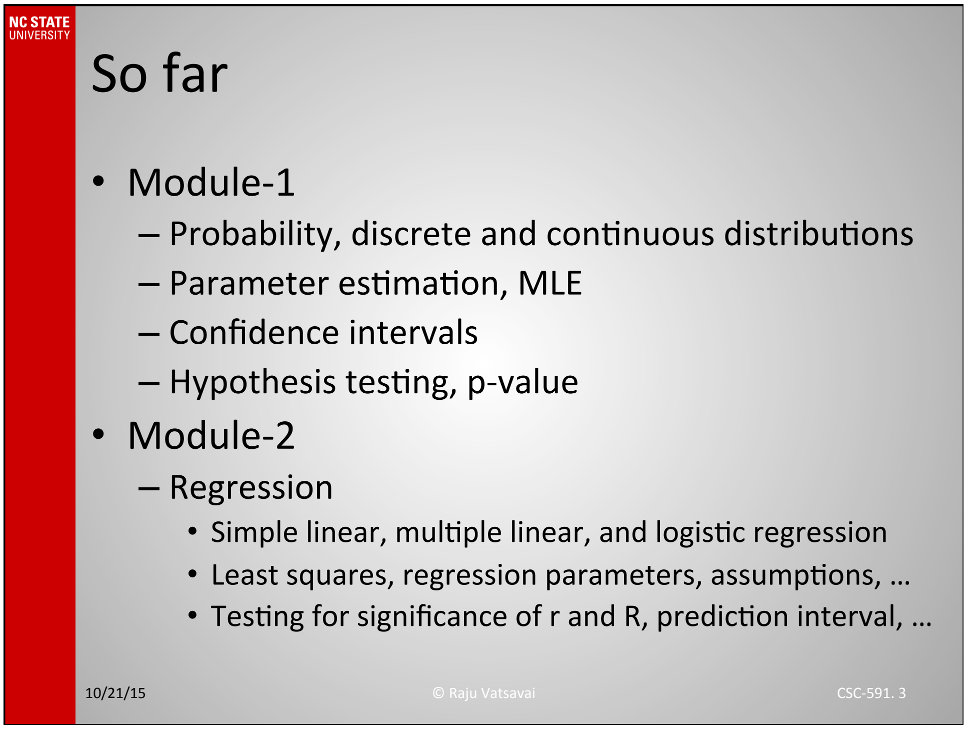
W10: 10/20/15-10/22/15



NC STATE UNIVERSITY

Admin: Changes in grading

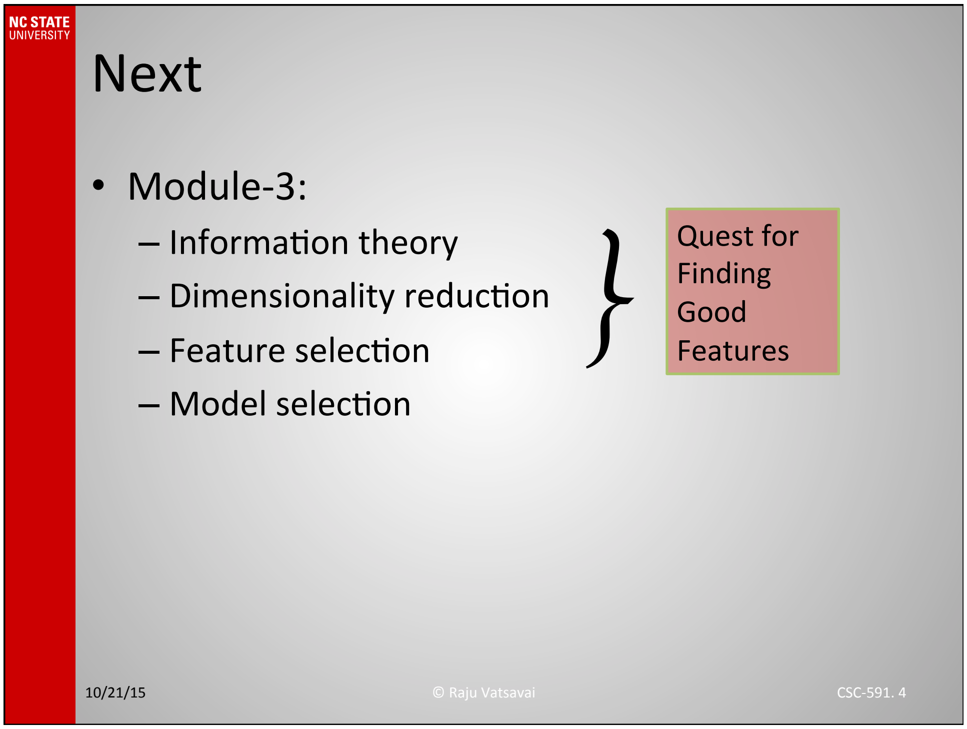
- Please send me an email before 10/23/15, if you want to keep 20% weightage to midterm-1.



NC STATE UNIVERSITY

So far

- Module-1
  - Probability, discrete and continuous distributions
  - Parameter estimation, MLE
  - Confidence intervals
  - Hypothesis testing, p-value
- Module-2
  - Regression
    - Simple linear, multiple linear, and logistic regression
    - Least squares, regression parameters, assumptions, …
    - Testing for significance of r and R, prediction interval, …



NC STATE UNIVERSITY

Next

- Module-3:
  - Information theory
  - Dimensionality reduction
  - Feature selection
  - Model selection

} Quest for Finding Good Features

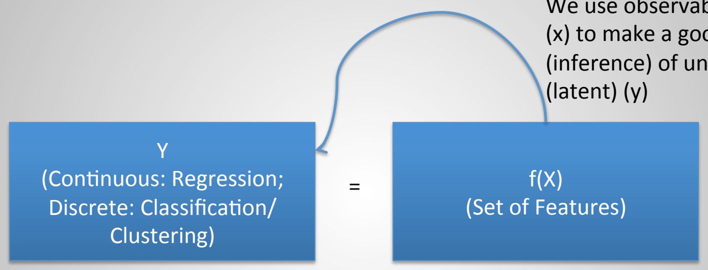
NC STATE UNIVERSITY

Today

- Information theory, entropy
- Readings
  - Chapter 2 from : [DM] David MacKay. Information Theory, Inference, and Learning Algorithms. (http://www.inference.phy.cam.ac.uk/itprnn/book.html)

NC STATE UNIVERSITY

Why do we need to select good features?

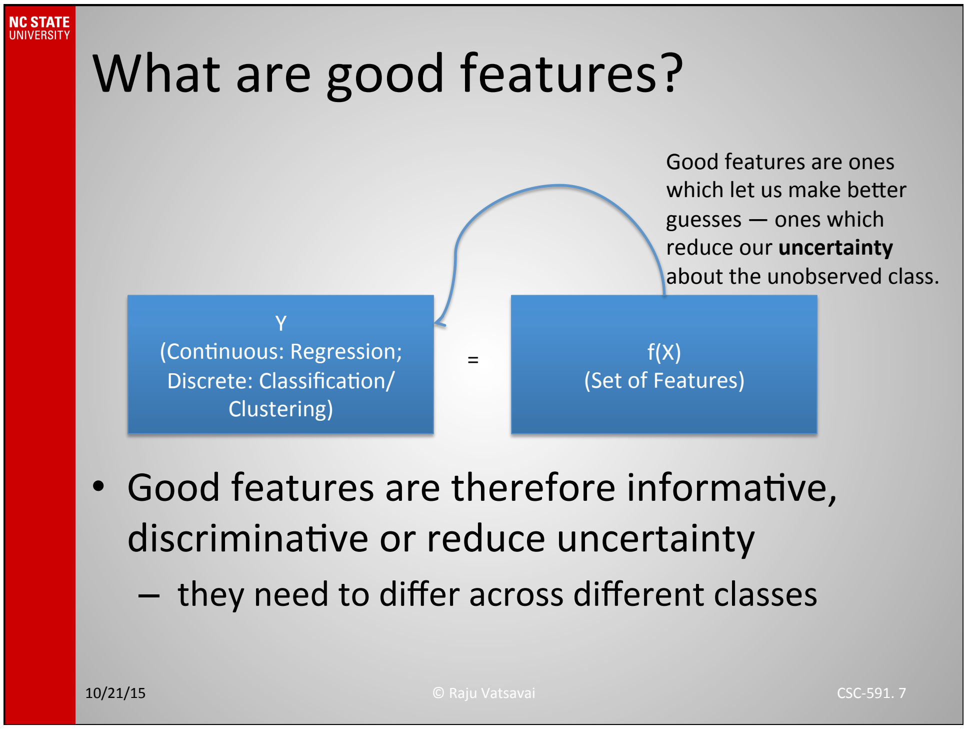

Y
(Continuous: Regression; Discrete: Classification/ Clustering)

=

f(X)
(Set of Features)

We use observable features (x) to make a good guess (inference) of unobservable (latent) (y)

- One can through all features in, and hope f() will figure-out the right ones
  - May be (regression, decision trees, …), but not always, inaddition “curse of dimensionality”



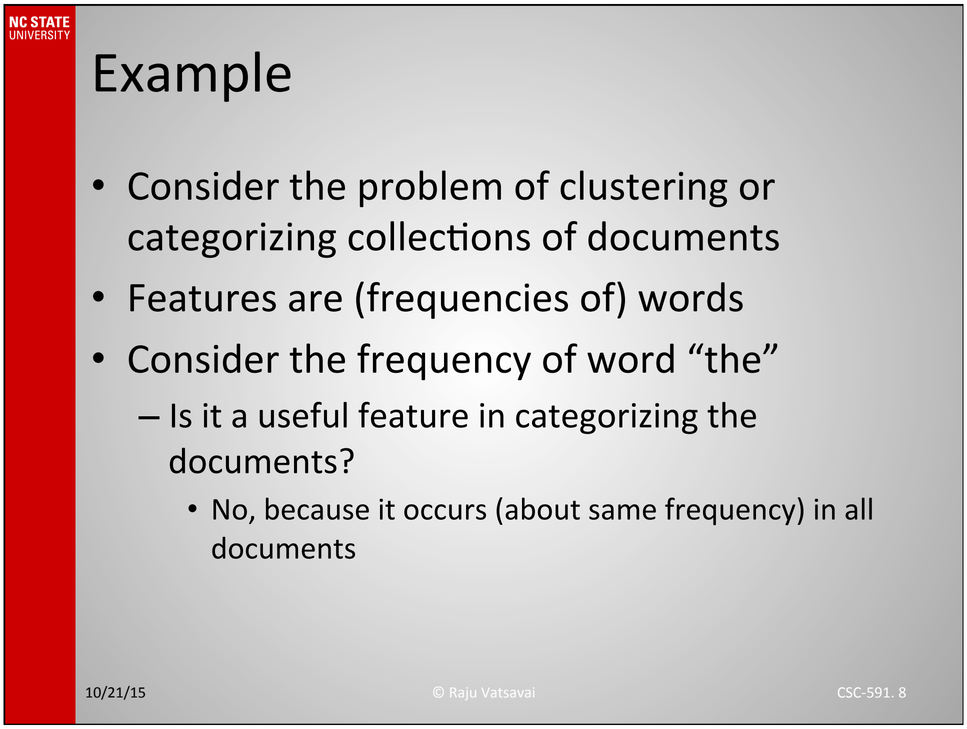
NC STATE UNIVERSITY

What are good features?

Good features are ones which let us make better guesses — ones which reduce our **uncertainty** about the unobserved class.

Y (Continuous: Regression; Discrete: Classification/ Clustering) = f(X) (Set of Features)

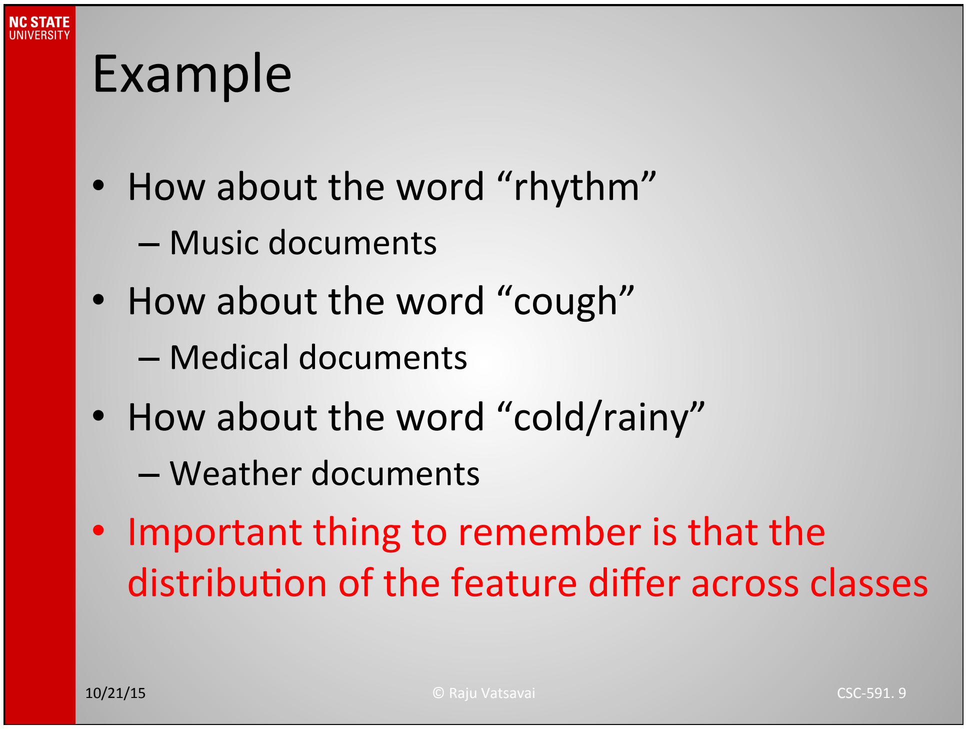
- Good features are therefore informative, discriminative or reduce uncertainty
  - they need to differ across different classes



NC STATE UNIVERSITY

Example

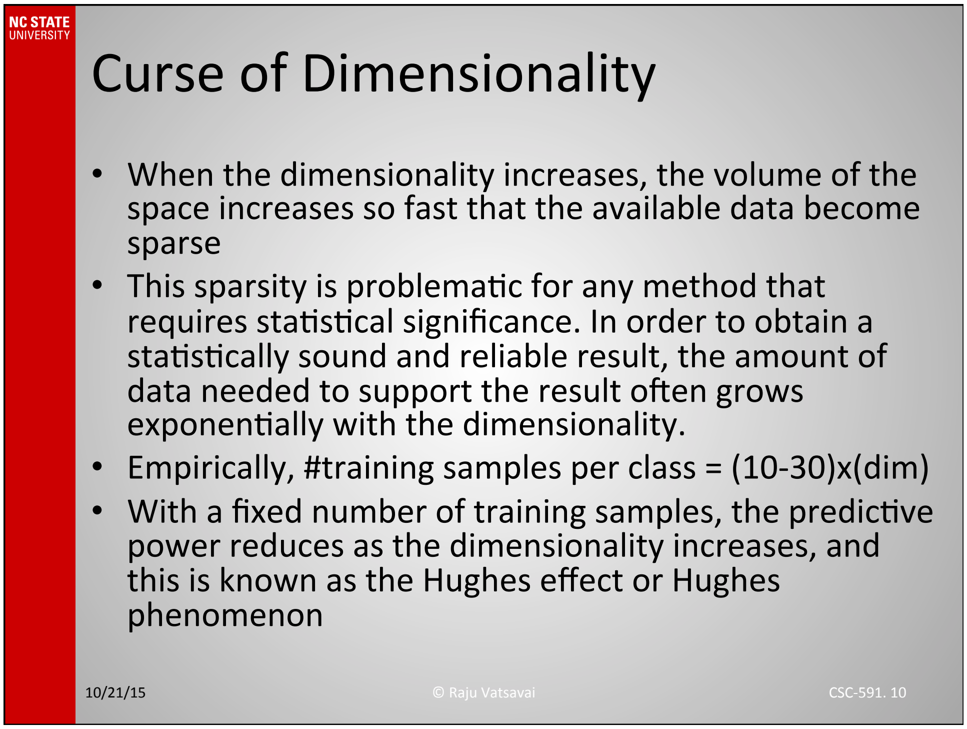
- Consider the problem of clustering or categorizing collections of documents
- Features are (frequencies of) words
- Consider the frequency of word "the"
  - Is it a useful feature in categorizing the documents?
    - No, because it occurs (about same frequency) in all documents



NC STATE UNIVERSITY

Example

- How about the word “rhythm”
  - Music documents
- How about the word “cough”
  - Medical documents
- How about the word “cold/rainy”
  - Weather documents
- Important thing to remember is that the distribution of the feature differ across classes



NC STATE UNIVERSITY

Curse of Dimensionality

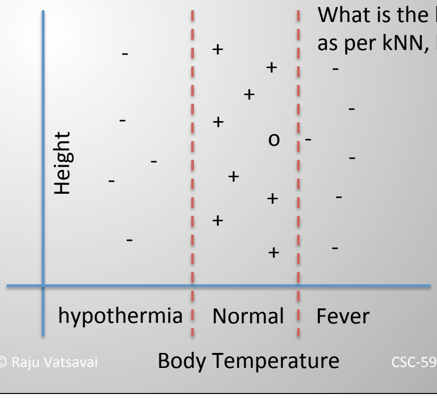
- When the dimensionality increases, the volume of the space increases so fast that the available data become sparse
- This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality.
- Empirically, #training samples per class = (10-30)x(dim)
- With a fixed number of training samples, the predictive power reduces as the dimensionality increases, and this is known as the Hughes effect or Hughes phenomenon

Irrelevant Attributes

- Irrelevant attributes do not contribute towards discriminating the object, but they do affect the geometric distance between objects (feature vectors)

If only body temperature is used, then $d(x,y) = \sqrt{(x_1 - y_1)^2} = |x_1 - y_1|$.

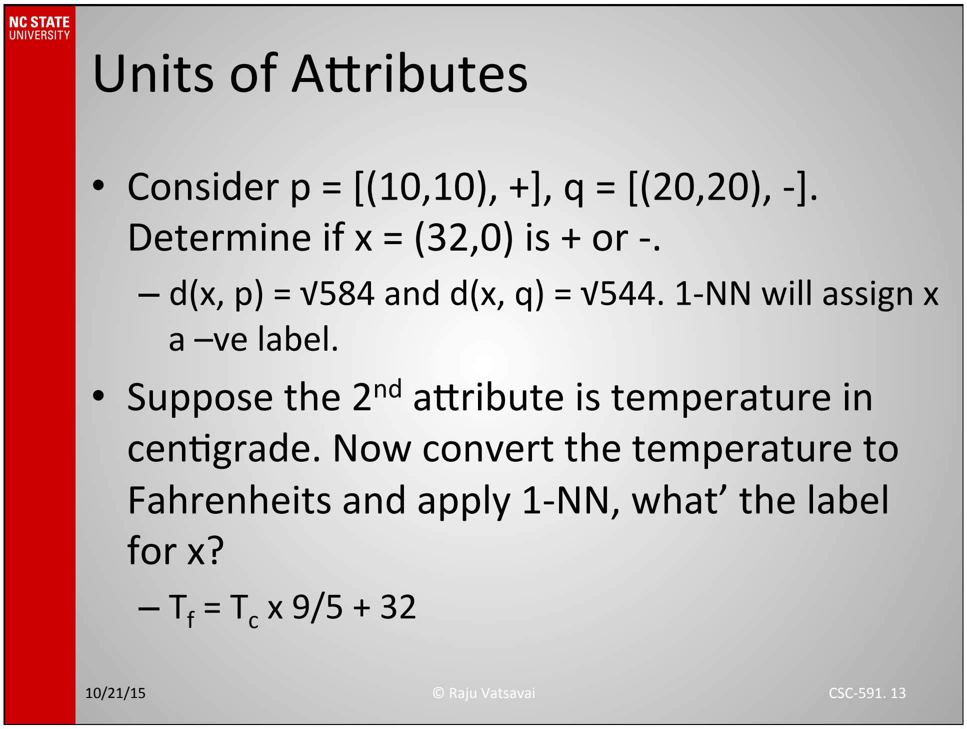
If both attributes are used, then $d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$. If the second attribute is irrelevant, then $(x_2 - y_2)^2$ is superfluous, and yet it affects adversely (e.g., kNN)



What is the label as per kNN, k=1

Scale of Attributes

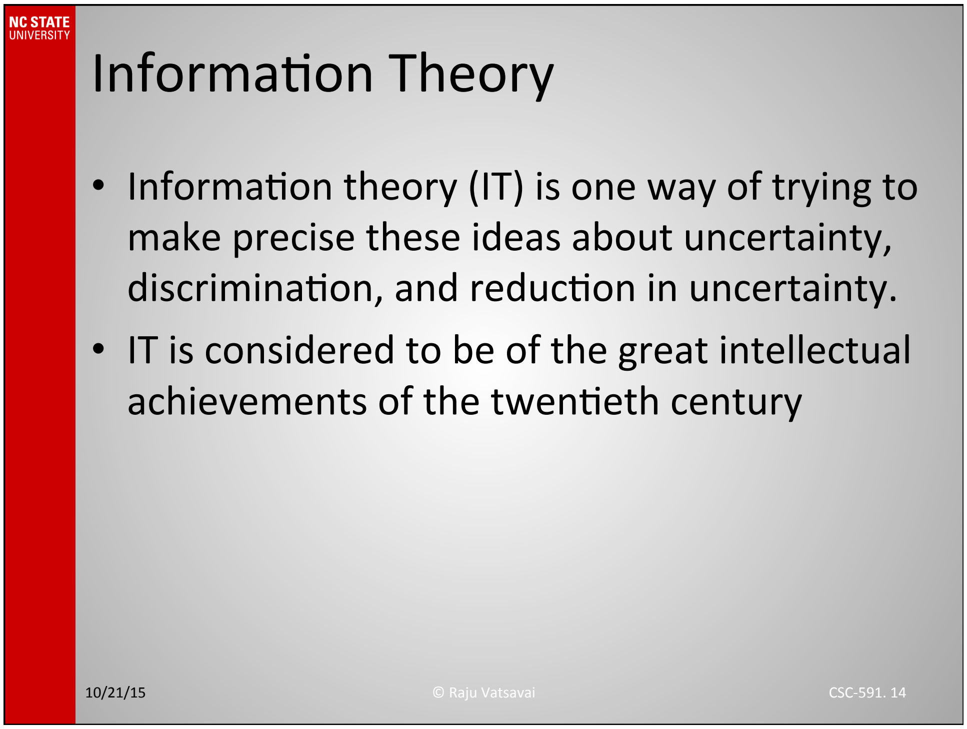
- Scale is another important aspect that impacts discrimination
- Consider two objects: x=(t,0.2,254) and y=(f, 0.1,194), 1st attribute is Boolean, 2nd is continuous with values in [0; 1]; and 3rd is continuous with values in [0; 1,000]. Then similarity is given by $d(x,y) = \sqrt{(1-0)^2+(0.2-0.1)^2+(254-194)^2}$. It is easy to see that 3rd attribute completely dominates, no matter what the values of 1st and 2nd attributes in their range of values.
  - Easy to correct this problem by scaling 3rd attribute (divide by 1000, which brings range to [0;1].



NC STATE UNIVERSITY

Units of Attributes

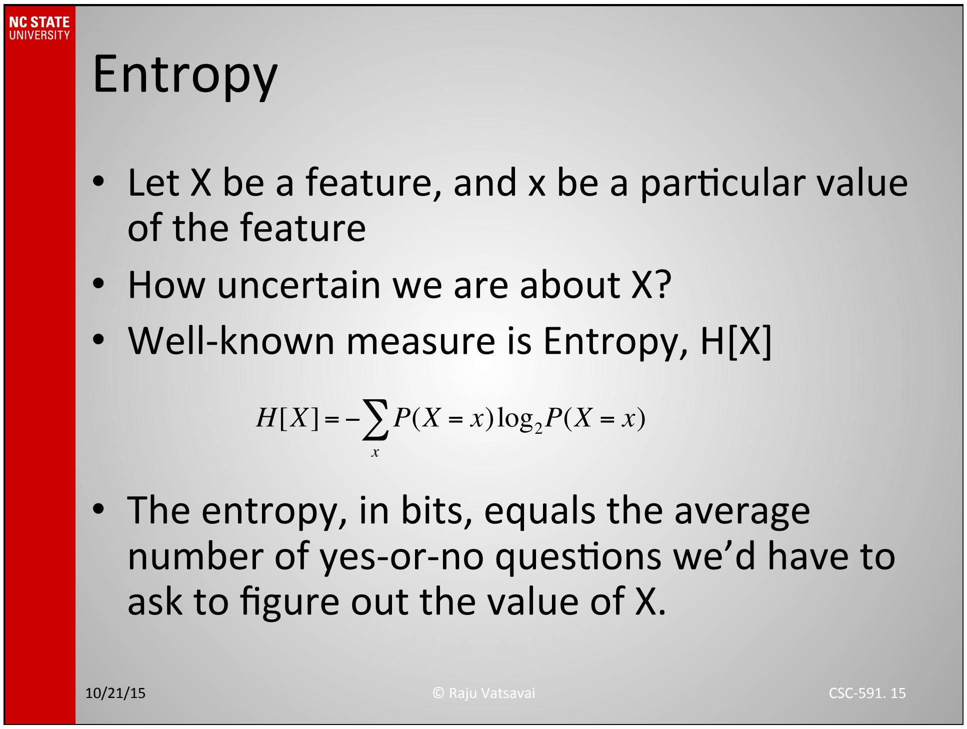
- Consider p = [(10,10), +], q = [(20,20), -]. Determine if x = (32,0) is + or -.
  - d(x, p) = √584 and d(x, q) = √544. 1-NN will assign x a –ve label.
- Suppose the 2nd attribute is temperature in centigrade. Now convert the temperature to Fahrenheits and apply 1-NN, what’ the label for x?
  - $T_f = T_c$ x 9/5 + 32



NC STATE UNIVERSITY

Information Theory

- Information theory (IT) is one way of trying to make precise these ideas about uncertainty, discrimination, and reduction in uncertainty.
- IT is considered to be of the great intellectual achievements of the twentieth century

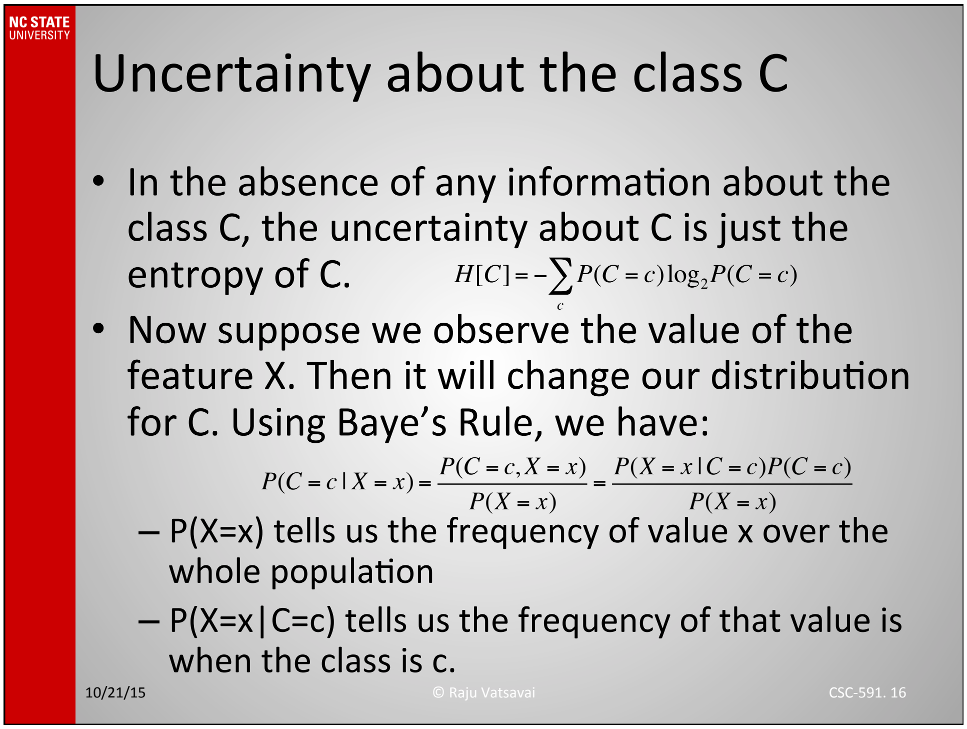


Entropy

- Let X be a feature, and x be a particular value of the feature
- How uncertain we are about X?
- Well-known measure is Entropy, H[X]

$$H[X] = -\sum_{x} P(X = x)\log_2 P(X = x)$$

- The entropy, in bits, equals the average number of yes-or-no questions we'd have to ask to figure out the value of X.

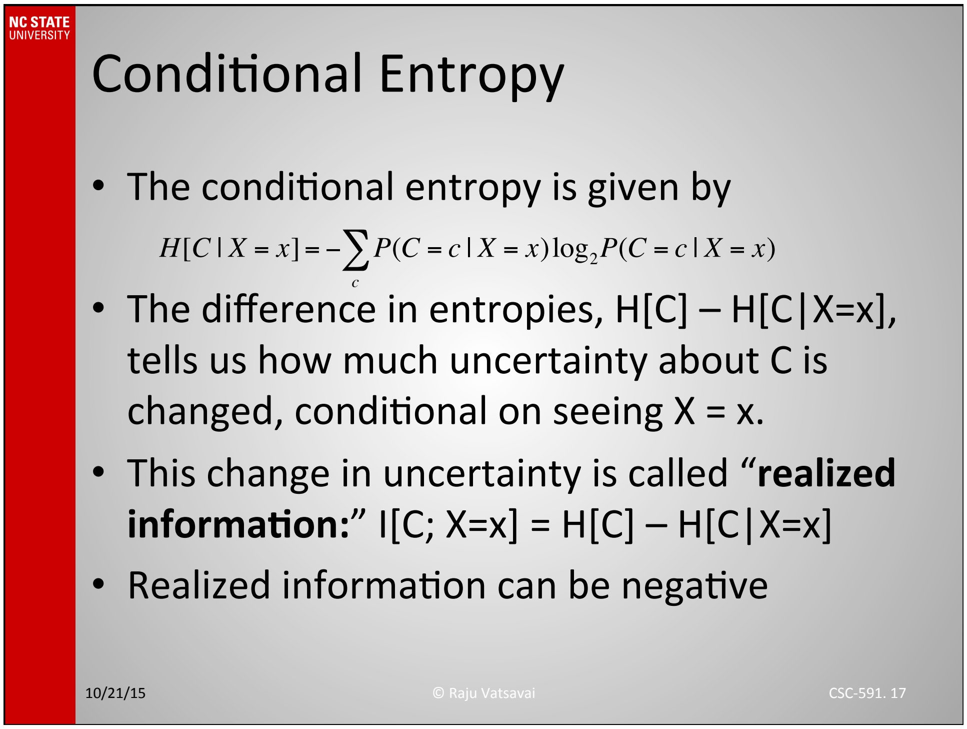


Uncertainty about the class C

- In the absence of any information about the class C, the uncertainty about C is just the entropy of C. $H[C] = -\sum_{c} P(C = c)\log_2 P(C = c)$
- Now suppose we observe the value of the feature X. Then it will change our distribution for C. Using Baye's Rule, we have:

$$P(C = c \mid X = x) = \frac{P(C = c, X = x)}{P(X = x)} = \frac{P(X = x \mid C = c)P(C = c)}{P(X = x)}$$

  - P(X=x) tells us the frequency of value x over the whole population
  - P(X=x|C=c) tells us the frequency of that value is when the class is c.



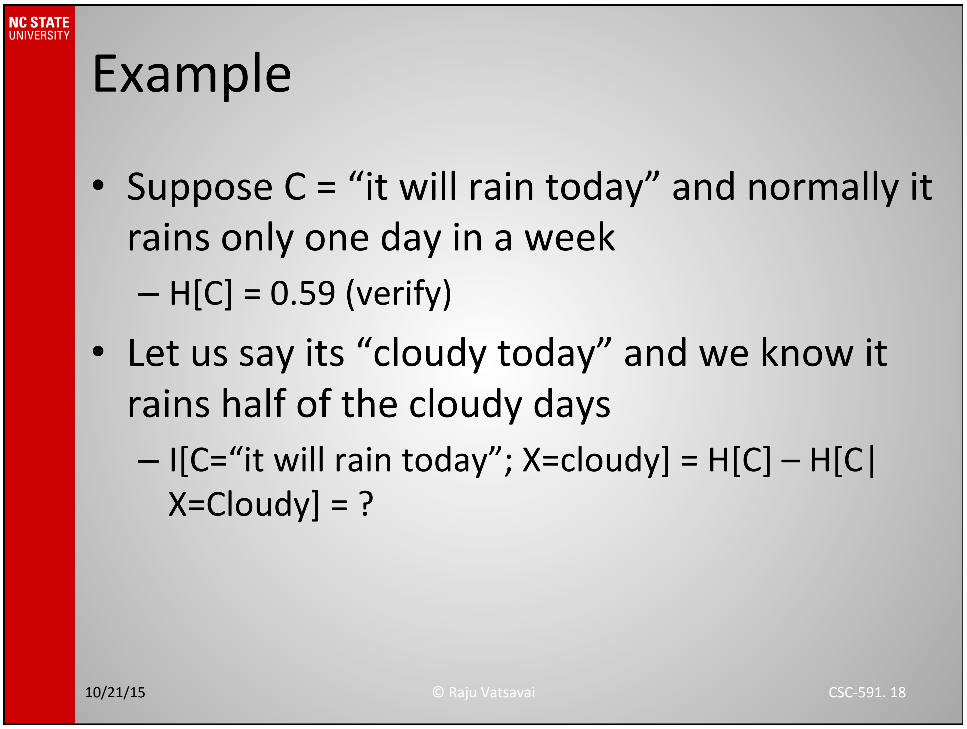
NC STATE UNIVERSITY

Conditional Entropy

- The conditional entropy is given by

$$H[C \mid X = x] = -\sum_{c} P(C = c \mid X = x)\log_2 P(C = c \mid X = x)$$

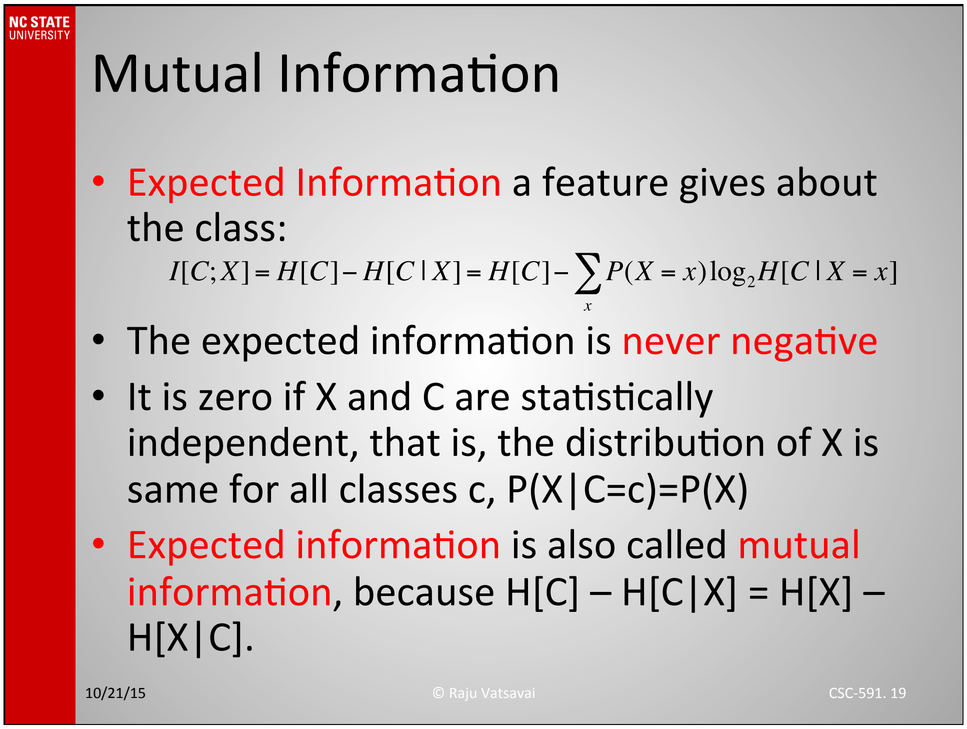
- The difference in entropies, H[C] – H[C|X=x], tells us how much uncertainty about C is changed, conditional on seeing X = x.
- This change in uncertainty is called "**realized information:**" I[C; X=x] = H[C] – H[C|X=x]
- Realized information can be negative



NC STATE UNIVERSITY

Example

- Suppose C = "it will rain today" and normally it rains only one day in a week
  - – H[C] = 0.59 (verify)
- Let us say its "cloudy today" and we know it rains half of the cloudy days
  - – I[C="it will rain today"; X=cloudy] = H[C] – H[C|X=Cloudy] = ?

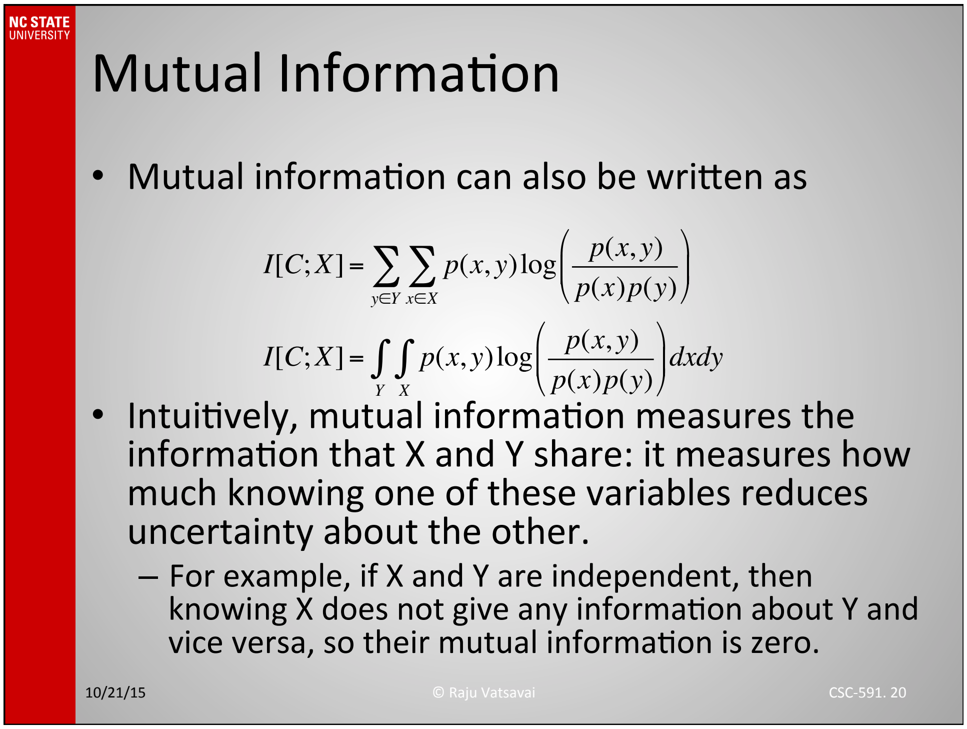


Mutual Information

- Expected Information a feature gives about the class:

$$I[C;X] = H[C] - H[C \mid X] = H[C] - \sum_{x} P(X = x) \log_2 H[C \mid X = x]$$

- The expected information is never negative
- It is zero if X and C are statistically independent, that is, the distribution of X is same for all classes c, P(X|C=c)=P(X)
- Expected information is also called mutual information, because H[C] – H[C|X] = H[X] – H[X|C].



Mutual Information

- Mutual information can also be written as

$$I[C;X] = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

$$I[C;X] = \int_{Y} \int_{X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$$

- Intuitively, mutual information measures the information that X and Y share: it measures how much knowing one of these variables reduces uncertainty about the other.
  - For example, if X and Y are independent, then knowing X does not give any information about Y and vice versa, so their mutual information is zero.

Entropy for Multiple Variables

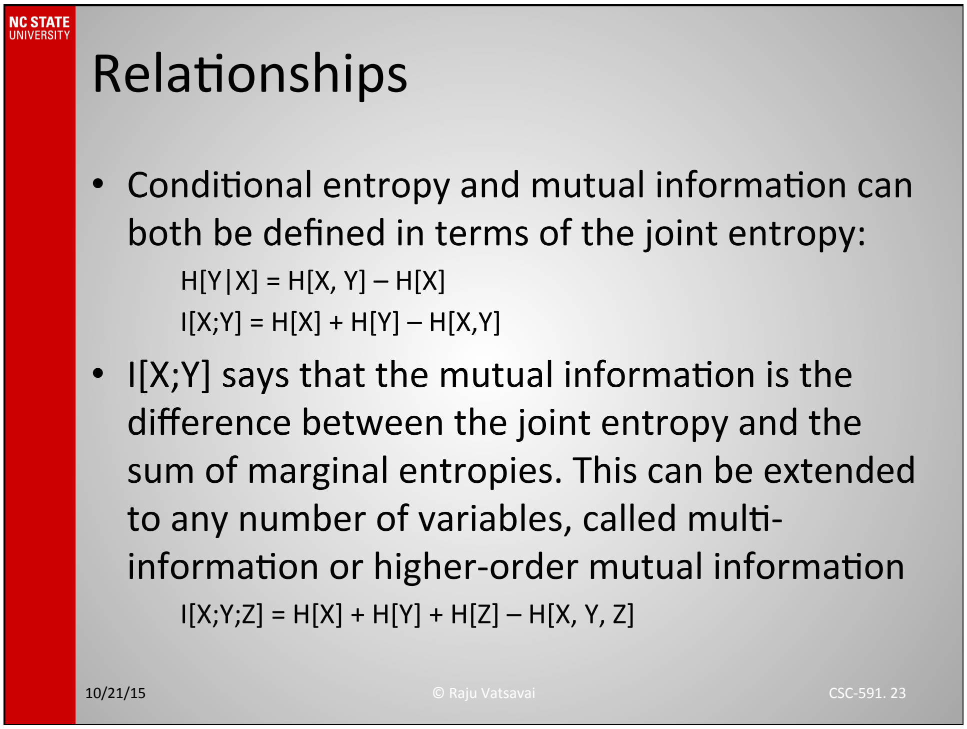
- Observation: There can be interactions among features, and that they can be more or less informative in different contexts
- The joint entropy of two random variables X and Y is just the entropy of their joint distribution

$$H[X,Y] \equiv -\sum_{x,y} P(X=x,Y=y)\log_2 P(X=x,Y=y)$$

  – This definition naturally extends to joint entropy of an arbitrary number of variables

Important properties

- A crucial property is that the joint entropy is sub-additive: H[X,Y] ≤ H[X] + H[Y], with equality if and only if X and Y are statistically independent
  – In terms of uncertainty, this says that you can't be more uncertain about the pair (X,Y) than you are about its components
  – In terms of coding, it says that the number of bits you would need to encode the pair is no more than the number of bits you would need to encode each of its members



NC STATE UNIVERSITY

Relationships

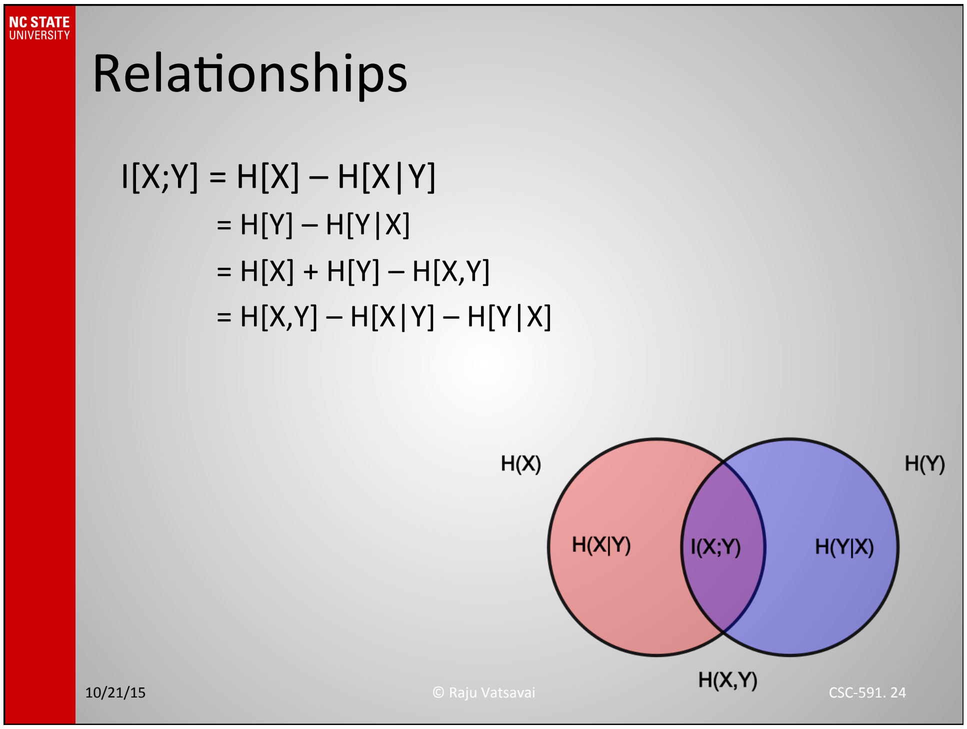
- Conditional entropy and mutual information can both be defined in terms of the joint entropy:

  H[Y|X] = H[X, Y] – H[X]

  I[X;Y] = H[X] + H[Y] – H[X,Y]

- I[X;Y] says that the mutual information is the difference between the joint entropy and the sum of marginal entropies. This can be extended to any number of variables, called multi-information or higher-order mutual information

  I[X;Y;Z] = H[X] + H[Y] + H[Z] – H[X, Y, Z]



NC STATE UNIVERSITY

Relationships

I[X;Y] = H[X] – H[X|Y]

= H[Y] – H[Y|X]

= H[X] + H[Y] – H[X,Y]

= H[X,Y] – H[X|Y] – H[Y|X]