CSDS 440: Machine Learning

Soumya Ray (he/him, sray@case.edu)

Olin 516

Office hours T, Th 11:15-11:45 or by appointment

Recap

- In supervised learning, examples are a___ by an o____.
- The “feature vector” representation creates a fixed size m____ where the rows are ____ and the columns are ____.
- Features can be n____, c____, o____ or h_____.
- What is the “feature space”?
- In the binary classification problem, the annotation is ____. This is called the c___ l___.
- What is the “decision boundary”?
- In a decision tree, internal nodes are a____ t____ and leaves are c___ l___.
- To classify a new examples we ______.
- Tree induction works through r____ p____. First we choose an a___ t___ if available. This creates d____ p____ from the data. We repeat until (1) ____ or (2) ______ happens.

Today

- Decision Tree Induction (Ch 3, Mitchell)
- Overfitting and overfitting control

Decision Tree Induction

- Given a set of examples, produce a decision tree
- Decision tree induction works using the idea of recursive partitioning
  - At each step, the algorithm will choose an attribute test
    - If no attribute looks good, return
  - The chosen test will partition the examples into disjoint partitions
  - The algorithm will then recursively call itself on each partition until
    - a partition only has data from one class (pure node) OR
    - it runs out of attributes

Choosing an Attribute

- Which attribute should we choose to test first?
  - Ideally, the one that is “most predictive” of the class label
    - i.e., the one that gives us the “most information” about what the label should be

- This idea is captured by the “(Shannon) entropy” of a random variable

Entropy of a Random Variable

- Suppose a random variable X has density $p(x)$. Its (Shannon) "entropy" is defined by:

$$H(X) = E(-\log_2(p(X)))$$

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

- Note: 0log(0) = 0 .

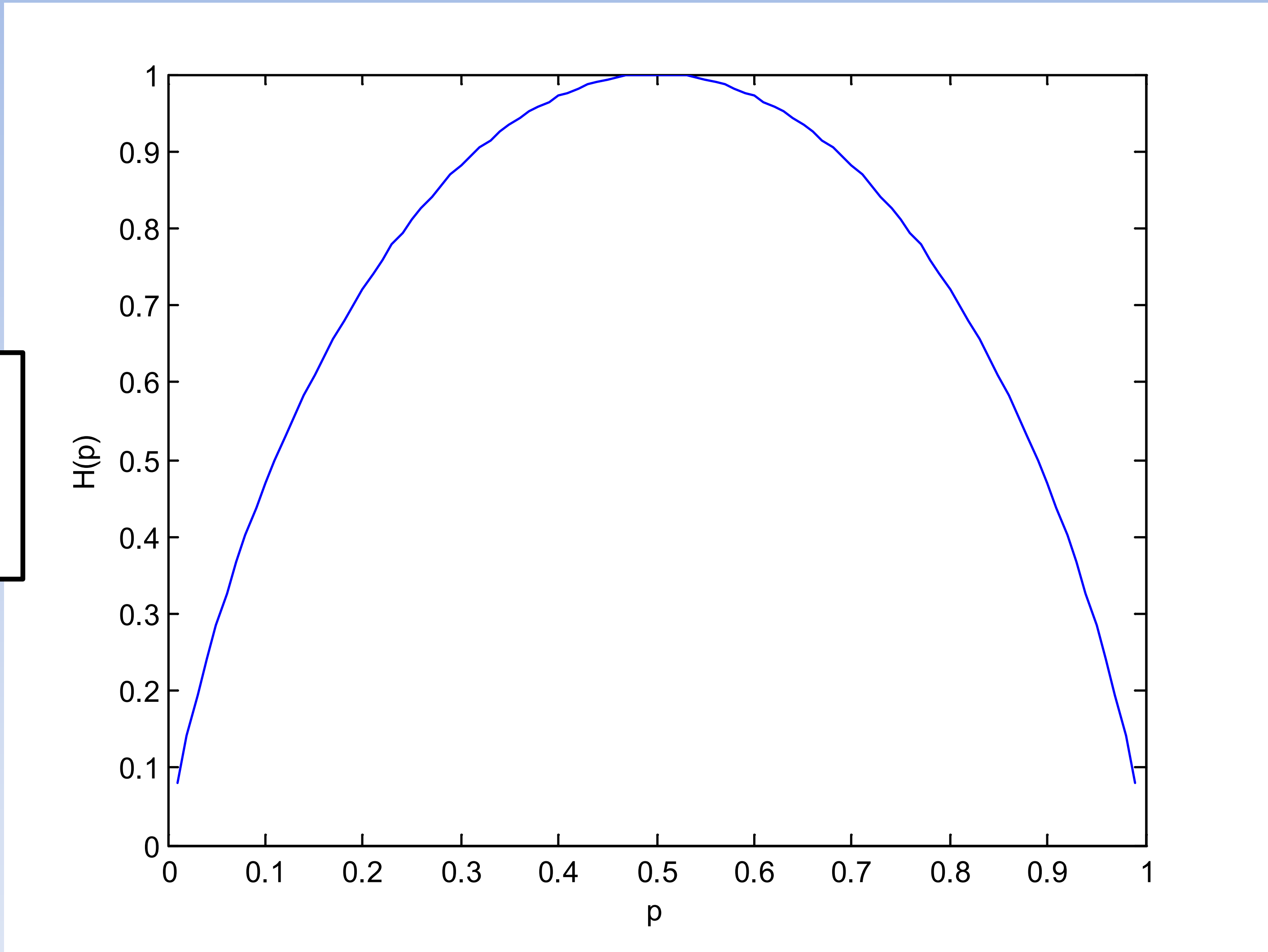
Example

- Suppose X has two values, 0 and 1, and pdf $p(0)=0.5, p(1)=0.5$
  - Then $H(X)=$?
- Suppose X has two values, 0 and 1, and pdf $p(0)=0.99, p(1)=0.01$
  - Then $H(X)=$?

0.081

- Suppose X has two values, 0 and 1, and pdf $p(0)=0.01, p(1)=0.99$
  - Then $H(X)=$?

Entropy of a Bernoulli r.v.



Entropy is typically denoted by *H(.)*

What is entropy?

- Measure of "information content" in a distribution
- Suppose we wanted to describe an r.v. X with n values and distribution $p(X=x)$
  - Shortest lossless description takes $-log_2(p(x))$ bits for each x
  - So entropy is the expected length of the shortest lossless description of the r.v.

Source Coding Theorem, Claude Shannon 1948

What's the connection?

- Entropy measures the *information content* of a random variable
- Suppose we treat the class variable, Y, as a random variable and measure its entropy
- Then we measure its entropy *after partitioning* the examples with an attribute X

The Entropy Connection

- The difference will be a measure of the "information gained" about Y by partitioning the examples with X

- So if we can choose the attribute X that maximizes this "information gain", we have found what we needed

The class as a random variable

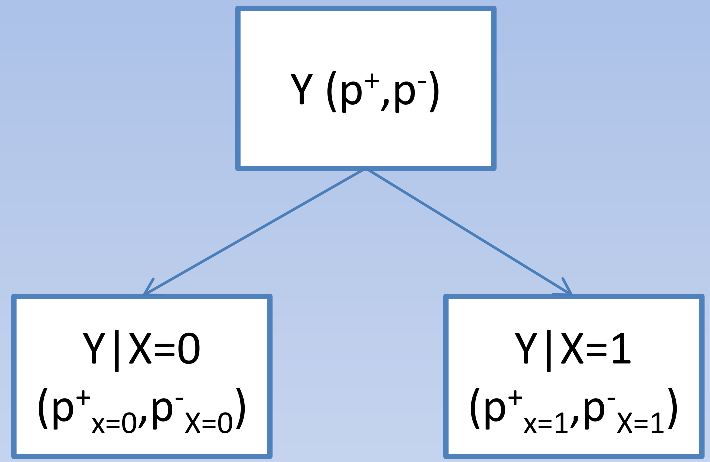
- Suppose at some point we have N training examples, of which pos are labeled "*positive*" and neg are labeled "*negative*" ($pos+neg=N$)
- We'll treat the class label as a Bernoulli r.v. Y that takes value 1 with prob. p^{+}=pos/N and 0 with prob. p^{-}=neg/N
- Then $H(Y)=-p^{+}log_2(p^{+})-p^{-}log_2(p^{-})$

Information Gain

- $IG(X)$=reduction in entropy of the class label if the data is partitioned using X

- Suppose an attribute X takes two values 1 and 0. After partitioning, we get the quantities $p^{+}_{X=1}, p^{-}_{X=1}, p^{+}_{X=0}$ and $p^{-}_{X=0}$. Then,

Information Gain contd.



$$H(Y \mid X=1) = -p^{+}_{X=1}\log_2 p^{+}_{X=1} - p^{-}_{X=1}\log_2 p^{-}_{X=1}$$

$$H(Y \mid X=0) = -p^{+}_{X=0}\log_2 p^{+}_{X=0} - p^{-}_{X=0}\log_2 p^{-}_{X=0}$$

$$H(Y \mid X) = p(X=1)H(Y \mid X=1) + p(X=0)H(Y \mid X=0)$$

$$IG(X) = H(Y) - H(Y \mid X)$$

Nominal Attributes

- If X has v values:

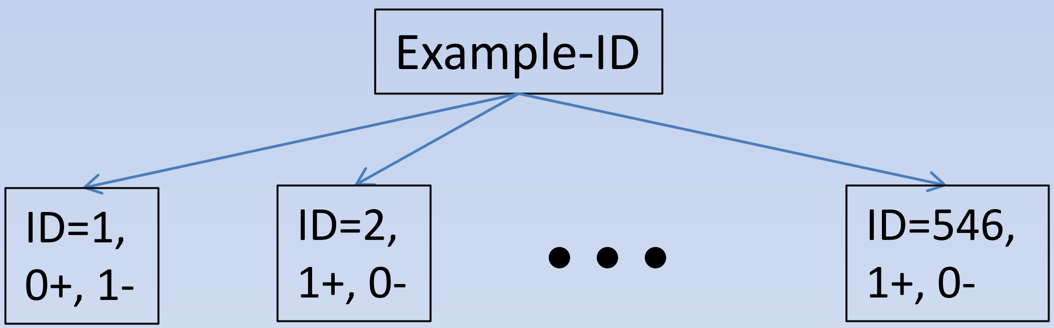
$$H(Y \mid X = v) = -p_{X=v}^{+} \log_2 p_{X=v}^{+} - p_{X=v}^{-} \log_2 p_{X=v}^{-}$$

$$H(Y \mid X) = \sum_{v} p(X = v) H(Y \mid X = v)$$

$$IG(X) = H(Y) - H(Y \mid X)$$

A Problem

- If an attribute has a lot of values, IG prefers it (resulting partitions tend to be pure)
- E.g., consider an “Example-ID” attribute



- This memorizes the data, so has perfect IG score

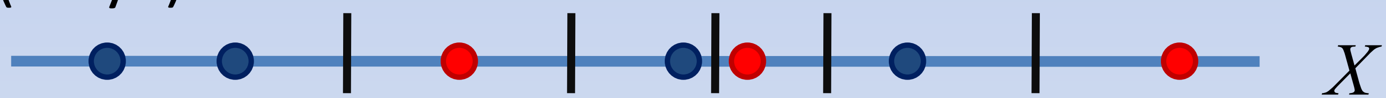
Fix: GainRatio

- Normalize IG with entropy of the attribute's distribution (computed from training data)

$$GR(X) = \frac{IG(X)}{H(X)}$$

Continuous Attributes

- Cannot test for equality
- Consider all Boolean tests of the form $X \geq v$ (or $X \leq v$)
  - Only values of interest are those v that separate adjacent training examples with different classes (why?)



- Note: In this case, the attribute cannot be removed, though the test ((attribute, value) tuple) can be