## Goal of the section today (4/28/2006)

Run through a concrete example of maximum entropy (maxent) models.
You should be able to understand these things at the end of the section:

- What are "features"
- What is being adjusted in the training process
- How to compute the objective function that's being optimized
- How to compute the derivative (used in optimization process)

---

This mini task is to classify animals to the category of cats, or bears.

$$c \in C = \{cat, bear\}$$

We have seen 3 animals. The first animal (d1) is fuzzy. It has claws and it's small.

$$d_1 = [fuzzy, claws, small]$$

We know it's a cat.

$$c_1 = cat$$

The second animal (d2) is fuzzy. It also has claws, but it's big.

$$d_2 = [fuzzy, claws, big]$$

We know it's a bear.

$$c_2 = bear$$

The third animal (d3) we've seen has claws, and its size is medium.

$$d_3 = [claws, medium]$$

We know it's a cat.

$$c_3 = cat$$

Question:

Here we have 5 characteristics that can be used to describe our data: being fuzzy, have claws, small size, big size, or medium size. And we have 2 classes: **cat** or **bear**.

How many (basic) feature functions do we have, and what are they?

## Feature Sets:

In this example, we have 10 features:

$f_1(\mathbf{c}, \mathbf{d}) = 1$ if **c** is cat and **d** is fuzzy
$f_2(\mathbf{c}, \mathbf{d}) = 1$ if **c** is bear and **d** is fuzzy
$f_3(\mathbf{c}, \mathbf{d}) = 1$ if **c** is cat and **d** has claws
$f_4(\mathbf{c}, \mathbf{d}) = 1$ if **c** is bear and **d** has claws
$f_5(\mathbf{c}, \mathbf{d}) = 1$ if **c** is cat and **d** is small
$f_6(\mathbf{c}, \mathbf{d}) = 1$ if **c** is bear and **d** is small
$f_7(\mathbf{c}, \mathbf{d}) = 1$ if **c** is cat and **d** is big
$f_8(\mathbf{c}, \mathbf{d}) = 1$ if **c** is bear and **d** is big
$f_9(\mathbf{c}, \mathbf{d}) = 1$ if **c** is cat and **d** is medium
$f_{10}(\mathbf{c}, \mathbf{d}) = 1$ if **c** is bear and **d** is medium

## Parameters:

We have 10 $\lambda_i$'s, each of them indicates how important each feature is.
Definition 1: $\text{vote}(\mathbf{c}) = \sum_i \lambda_i f_i(\mathbf{c},\mathbf{d})$
In our example…
Suppose we already have a set of $\lambda_i$'s. (see the tables below)
For the first animal $d_1$ = [fuzzy, claws, small]
$\text{vote}(\mathbf{cat}) = \sum_{i=1to10} \lambda_i f_i(\mathbf{cat},d_1) =$ **-0.2**

| | | | | | |
|---|---|---|---|---|---|
| $\lambda_1=$ | -1 | $f_1(\mathbf{cat},d_1) =$ | 1 | $\lambda_1\ f_1(\mathbf{cat},d_1) =$ | -1 |
| $\lambda_2=$ | 1 | $f_2(\mathbf{cat},d_1) =$ | 0 | $\lambda_2\ f_2(\mathbf{cat},d_1) =$ | 0 |
| $\lambda_3=$ | 0.5 | $f_3(\mathbf{cat},d_1) =$ | 1 | $\lambda_3\ f_3(\mathbf{cat},d_1) =$ | 0.5 |
| $\lambda_4=$ | -0.5 | $f_4(\mathbf{cat},d_1) =$ | 0 | $\lambda_4\ f_4(\mathbf{cat},d_1) =$ | 0 |
| $\lambda_5=$ | 0.3 | $f_5(\mathbf{cat},d_1) =$ | 1 | $\lambda_5\ f_5(\mathbf{cat},d_1) =$ | 0.3 |
| $\lambda_6=$ | -0.3 | $f_6(\mathbf{cat},d_1) =$ | 0 | $\lambda_6\ f_6(\mathbf{cat},d_1) =$ | 0 |
| $\lambda_7=$ | -0.6 | $f_7(\mathbf{cat},d_1) =$ | 0 | $\lambda_7\ f_7(\mathbf{cat},d_1) =$ | 0 |
| $\lambda_8=$ | 0.6 | $f_8(\mathbf{cat},d_1) =$ | 0 | $\lambda_8\ f_8(\mathbf{cat},d_1) =$ | 0 |
| $\lambda_9=$ | 0.8 | $f_9(\mathbf{cat},d_1) =$ | 0 | $\lambda_9\ f_9(\mathbf{cat},d_1) =$ | 0 |
| $\lambda_{10}=$ | -0.8 | $f_{10}(\mathbf{cat},d_1) =$ | 0 | $\lambda_{10}\ f_{10}(\mathbf{cat},d_1) =$ | 0 |
| | | | | vote(**cat**)= | **-0.2** |

The vote for the other class, bear, is:

vote(**bear**) = $\sum_{i=1 to 10} \lambda_i f_i$(**bear**, $d_1$) = **0.2**

| | | | | | |
|---|---|---|---|---|---|
| $\lambda_1$= | -1 | $f_1$(**bear**,$d_1$) = | 0 | $\lambda_1$ $f_1$(**bear**,$d_1$) = | 0 |
| $\lambda_2$= | 1 | $f_2$(**bear**,$d_1$) = | 1 | $\lambda_2$ $f_2$(**bear**,$d_1$) = | 1 |
| $\lambda_3$= | 0.5 | $f_3$(**bear**,$d_1$) = | 0 | $\lambda_3$ $f_3$(**bear**,$d_1$) = | 0 |
| $\lambda_4$= | -0.5 | $f_4$(**bear**,$d_1$) = | 1 | $\lambda_4$ $f_4$(**bear**,$d_1$) = | -0.5 |
| $\lambda_5$= | 0.3 | $f_5$(**bear**,$d_1$) = | 0 | $\lambda_5$ $f_5$(**bear**,$d_1$) = | 0 |
| $\lambda_6$= | -0.3 | $f_6$(**bear**,$d_1$) = | 1 | $\lambda_6$ $f_6$(**bear**,$d_1$) = | -0.3 |
| $\lambda_7$= | -0.6 | $f_7$(**bear**,$d_1$) = | 0 | $\lambda_7$ $f_7$(**bear**,$d_1$) = | 0 |
| $\lambda_8$= | 0.6 | $f_8$(**bear**,$d_1$) = | 0 | $\lambda_8$ $f_8$(**bear**,$d_1$) = | 0 |
| $\lambda_9$= | 0.8 | $f_9$(**bear**,$d_1$) = | 0 | $\lambda_9$ $f_9$(**bear**,$d_1$) = | 0 |
| $\lambda_{10}$= | -0.8 | $f_{10}$(**bear**,$d_1$) = | 0 | $\lambda_{10}$ $f_{10}$(**bear**,$d_1$) = | 0 |
| | | | | vote(**bear**)= | **0.2** |

Definition 2: probabilistic model

$$P(\mathbf{c} \mid \mathbf{d}, \lambda) = \frac{\exp \sum_i \lambda_i f_i(\mathbf{c},\mathbf{d})}{\sum_{c'} \exp \sum_i \lambda_i f_i(\mathbf{c'},\mathbf{d})} = \frac{\exp(\text{vote}(\mathbf{c}))}{\sum_{c'} \exp(\text{vote}(\mathbf{c'}))}$$

In our example…

$$P(\mathbf{cat} \mid d_1, \lambda) = \frac{\exp(\text{vote}(\mathbf{cat}))}{\exp(\text{vote}(\mathbf{cat})) + \exp(\text{vote}(\mathbf{bear}))} = \frac{\exp(-0.2)}{\exp(-0.2) + \exp(0.2)} = \mathbf{0.4013}$$

$$P(\mathbf{bear} \mid d_1, \lambda) = \frac{\exp(\text{vote}(\mathbf{bear}))}{\exp(\text{vote}(\mathbf{cat})) + \exp(\text{vote}(\mathbf{bear}))} = \frac{\exp(0.2)}{\exp(-0.2) + \exp(0.2)} = \mathbf{0.5987}$$

Interpretation from this example:

Given the set of $\lambda_i$'s in the table, and given that we see an animal with the features [fuzzy, claws, small], we'll conclude the probability of it being a cat is **0.4013**, being a bear is **0.5987**. So we'll say it's a bear.

If we go back to our first page, we'll see that this animal is in our training data, and it's actually a cat, not a bear!

Question: Intuitively, how do we adjust the $\lambda_i$'s so that we can correctly predict this example?

## What are we optimizing?

When we're adjusting the $\lambda_i$'s, we're aiming at maximizing the (conditional) likelihood of our training data.

$$P(C \mid D, \lambda) = \prod_{(c,d)\in(C,D)} P(c \mid d, \lambda)$$

It's equivalent to maximizing the log conditional likelihood.

$$\log P(C \mid D, \lambda) = \sum_{(c,d)\in(C,D)} \log P(c \mid d, \lambda)$$

Sum over all training examples

In Definition 2

## What's necessary for doing the optimization?

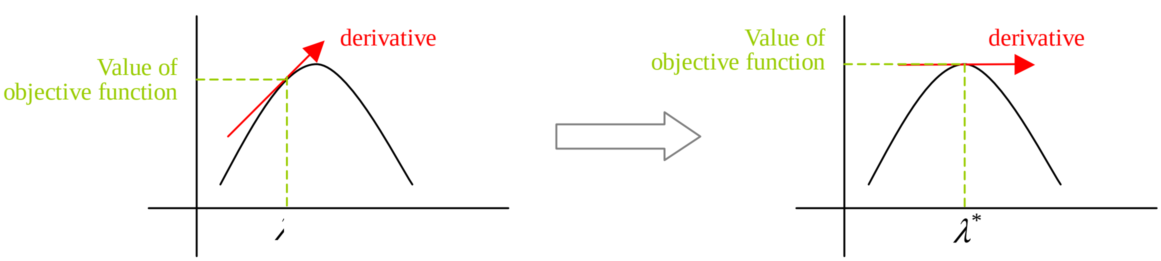
Give a set of $\lambda_i$'s, calculate

1. Objective : the conditional likelihood of the data → $\log P(C \mid D, \lambda)$
2. Derivatives :

$$\frac{\partial \log P(C \mid D, \lambda)}{\partial \lambda_i} = \text{actual count}(f_i, \text{C}) - \text{predicted count}(f_i, \lambda)$$

$$= \sum_{(c,d)\in(C,D)} f_i(c,d) - \sum_{(c,d)\in(C,D)} \sum_{c'} P(c' \mid d, \lambda)\, f_i(c',d)$$

A simple intuition here: (in one-dimensional space):




See the excel file for a detailed example of how to compute the value of the objective function and derivatives, and how to adjust $\lambda_i$'s.