Conditional Maxent Models for Classification

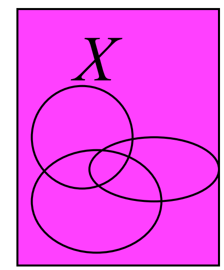
The relationship between conditional and joint maxent/exponential models

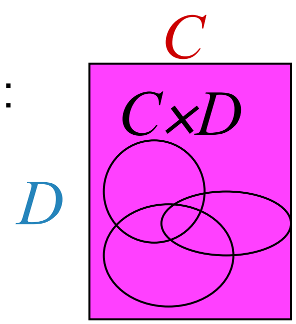
Classification

- What do these joint models of $\mathrm{P}(X)$ have to do with conditional models $\mathrm{P}(C|D)$?
- Think of the space $C \times D$ as a complex X.
  - C is generally small (e.g., 2-100 topic classes)
  - D is generally huge (e.g., space of documents)
- We can, in principle, build models over $\mathrm{P}(C,D)$.
- This will involve calculating expectations of features (over $C \times D$):

$$E(f_i) = \sum_{(c,d) \in (C,D)} P(c,d) f_i(c,d)$$

- Generally impractical: can't enumerate X efficiently.





Christopher Manning

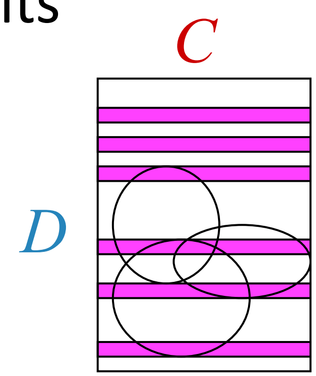
Classification II

- *D* may be huge or infinite, but only a few *d* occur in our data.
- What if we add one feature for each *d* and constrain its expectation to match our empirical data?

$$\forall (d) \in D \quad P(d) = \hat{P}(d)$$

- Now, most entries of *P*(*c*,*d*) will be zero.
- We can therefore use the much easier sum:

$$E(f_i) = \sum_{(c,d)\in(C,D)} P(c,d) f_i(c,d)$$
$$= \sum_{(c,d)\in(C,D)\wedge \hat{P}(d)>0} P(c,d) f_i(c,d)$$



Classification III

- But if we've constrained the *D* marginals

$$\forall(d) \in D \quad P(d) = \hat{P}(d)$$

- then the only thing that can vary is the conditional distributions:

$$P(c,d) = P(c \mid d)P(d)$$
$$= P(c \mid d)\hat{P}(d)$$

Classification IV

- This is the connection between joint and conditional maxent / exponential models:
  - Conditional models can be thought of as joint models with marginal constraints.
- Maximizing joint likelihood and conditional likelihood of the data in this model are equivalent!

Conditional Maxent Models for Classification

The relationship between conditional and joint maxent/ exponential models