A Virtual Manipulative for Learning Log-Linear Models

Author 1
XYZ Company
111 Anywhere Street
Mytown, NY 10000, USA
`author1@xyz.org`

Author 2
XYZ Company
111 Anywhere Street
Mytown, NY 10000, USA
`author1@xyz.org`

Abstract

Abstract here.

1 Introduction

except for reading of data files, purely client-side $\Rightarrow$ very easy to set-up; open-source; data input format makes it extensible; individual lessons can be tailored (e.g., hide/show buttons, different tooltips for lessons)

2 Model

Our aim is to teach an intuitive understanding conditional log-linear models. Given N data points $\{(x_i, y_i)\}_{i=1}^N$, we are interested in estimating distributions

$$\hat{p}(y \mid x) = \frac{u(x,y)}{\sum_{y'} u(x,y')}, \tag{1}$$

where $u(x,y)$ represents an unnormalized probability

$$\begin{aligned} u(x,y) &= \exp\left(\vec{\theta} \cdot \vec{f}(x,y)\right) \qquad (2) \\ &= \exp\left(\sum_{k=1}^{K} \theta_k f_k(x,y)\right). \qquad (3) \end{aligned}$$

3 Our Notes

[FF: These are simply copied from the Google doc titled “600.465: Maxent Notes.” This section could be retitled general pedagogical aims, or something of the sort.]

- If the striped feature is predicted to occur less often than it actually does, you should raise its weight.
- Its possible to overfit the training data. Regularization compensates for that and can in fact make you underfit.
  - In particular, weights may zoom off to +infinity or -infinity if a feature is always or never present on the *observed* examples (may need to cook special datasets for this)
- Interactions:
  - Raising one weight may reduce or reverse the need to raise other weights. This can be seen by watching the gradient as we slide the slider.
  - Can share features across conditions and this helps regularizer even if likelihood is the same
  - Features that only fire on conditions have no effect on conditional distribution
  - Feature conjunctions: fewer vs. more features
  - Feature that everything/nothing has — weights go to $\pm\infty$
  - Opposing features, e.g., solid vs striped, where there are only 2 options (or, red vs. blue)
- Likelihood always goes up if you follow gradient
  - gradient = observed - expected count (-regularizer)
  - This is evident in the LL-bar at the top
- LL is maximized when you match the empirical (except for overfitting?)

- Frequent conditions more influential
- Some distributions cant be matched — but you get generalization
- The initial setting where all weights = 0 gives the uniform distribution (in each condition).
  - Some further understanding of the entropy view? (See below.)

4 Usability

“New Counts” button The other use is to help the user experiment with datasets of different sizes, by changing N to scale the counts and then clicking ”New counts.”