Iterative Scaling Algorithm

Given:
- Training corpus, \((\vec{x}_j, c_j)\)
- Feature values for \(f_i\) on training corpus
- Empirical distribution \(\tilde{p}\)

Assumptions:
- All features are binary
- For all \(j\), \(\sum_i f_i(\vec{x}_j, c_j) = k\)

1. Initialise \(\lambda_i := 0\) for each feature \(i\)
2. Compute \(E_{\tilde{p}}(f_i)\) for each feature \(i\)
3. Compute \(p_{\lambda}(\vec{x}_j, c_j)\) for each datum \(j\)
4. Compute \(E_{p_{\lambda}}(f_i)\), for each feature \(i\)
5. For all \(i\):
   \[
   \lambda_i := \lambda_i + \frac{1}{k}(\log \frac{E_{\tilde{p}}(f_i)}{E_{p_{\lambda}}(f_i)})
   \]
6. Goto (3) if not converged
Iterative Scaling Algorithm

Advantages:

- incorporates feature selection,
- scales up well in numbers of features w.r.t. other methods,
- resilient to feature dependence.

Disadvantages:

- slow in training relative to other classification methods,
- problems with smoothing,
- binary and summation assumptions (not a big deal).