Basics

Random variable x ranges over possible events, including assigned classes

- Probability distribution p tells us the class of a specific data point—e.g., $P(\text{document}_1) = \text{positive sentiment}$
- Evidence is encoded as k features
  - Assume binary features for simplicity, i.e., $f_j(x) \in \{0, 1\}$
  - Expected value of feature f_j, where $j \in [1 \ldots k]$ (p is actual)

$$E_p f_j = \sum_x p(x) f_j(x)$$

- Expected value based on the evidence (training sample), $\bar{p}$

$$E_{\bar{p}} f_j = \sum_x \bar{p}(x) f_j(x)$$

- Our learned probability distribution, p, must respect the evidence (training sample), $\bar{p}$

$$E_p f_j = E_{\bar{p}} f_j$$

- Many possible solutions for p (actual) given $\bar{p}$ (evidence)

Principle of Maximum Entropy

Discussion follows Ratnaparkhi

Choose the hypothesis that has the maximum entropy because it makes the least unjustified assumptions

- Entropy (Shannon, inventor of information theory)

$$H(p) = -\sum_{x} p(x) \log p(x)$$

  - Indicates lack of knowledge or "disorder"
  - The greater the entropy of a distribution the more information you need to describe it
- $P = \{p | \forall j\colon E_p f_j = E_{\bar{p}} f_j\}$
  - P is the set of distributions that match the evidence
- MaxEnt: Given evidence, $\bar{p}$, choose p^* where

$$p^* = \operatorname*{argmax}_{p \in P} H(p)$$

Relative entropy: Kullback-Leibler Divergence

- Interpret q as an approximation to a true distribution p
- Then, $D(p, q)$ is the information to be added to q to make it equal p

$$D(p,q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} = \sum_{x} p(x) \log p(x) - \sum_{x} p(x) \log q(x)$$

  - Use $0 \log 0 = 0$
  - $D(p, q) \geq 0$
  - $D(p, q) = 0$ if and only if $p = q$
- Sometimes written $D(p \parallel q)$
- Not a distance since it is not symmetric
- Symmetric variants, e.g., Jensen-Shannon Divergence, whose positive square root is the Jensen-Shannon Distance

Example of Kullback-Leibler Divergence

Not symmetric

- Consider distributions over three colors: red, green, blue
- Let $p = [\frac{1}{2}, \frac{1}{2}, 0]$
- Let $q = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$
- $D(p, q) = ?$
- $D(q, p) = ?$

Definitions

- The α_j are parameters
- Q is the set of distributions of the form of the product of positive constants exponentiated by feature functions (π to normalize)

$$Q = \{p | p(x) = \pi \prod_j \alpha_j^{f_j(x)}\}, \text{ where } \alpha_j \in (0, \infty)$$

- Let $t \in Q$

$$\log t(x) = \log \pi + \sum_j f_j(x) \log \alpha_j$$

Distributions in P are Interchangeable Relative to Q

Recall P are evidence-matching and Q are feature-based distributions

- Let $r, s \in P$, $t \in Q$

$$\sum_x r(x) \log t(x) = \sum_x r(x)[\log \pi + \sum_j f_j(x) \log \alpha_j]: \text{ expansion}$$

$$= \log \pi[\sum_x r(x)] + \sum_j \log \alpha_j \sum_x r(x) f_j(x)]: \text{ grouping}$$

$$= \log \pi[\sum_x s(x)] + \sum_j \log \alpha_j \sum_x s(x) f_j(x)]: \text{ } r,\ s \text{ sum to 1; } r, s \in P$$

$$= \sum_x s(x) \log t(x): \text{ reverse of above step}$$

- Used $\sum_x r(x) = 1 = \sum_x s(x)$ above

Pythagorean Property

- Let $p \in P$ (evidence matching), $q \in Q$ (features as exponentials), and $p^* \in P \cap Q$ (both)—omitting (x) for brevity
$$D(p,q) = D(p,p^*) + D(p^*,q)$$
- Begin from right-hand side
$$D(p,p^*) + D(p^*,q)$$

$$= \sum_x p \log p - \sum_x p \log p^* + \sum_x p^* \log p^* - \sum_x p^* \log q\text{: definition}$$

$$= \sum_x p \log p - \sum_x p \log p^* + \sum_x p \log p^* - \sum_x p \log q\text{: previous page}$$

$$= \sum_x p \log p - \sum_x p \log q\text{: middle two terms cancel out}$$

$$= D(p,q)$$

Maximum Entropy Solution: Existence

$$\text{If } p^* \in P \cap Q, \text{ then } p^* = \operatorname*{argmax}_{p \in P} H(p)$$

- Let $p \in P$, $p^* \in P \cap Q$, $u \in Q$, where u is the uniform distribution
  - $u \in Q$ because we can set each $\alpha_j = 1$
  - For X possibilities, $u(x) = \frac{1}{|X|}$

$$D(p,u) = \sum_x p(x) \log p(x) - \sum_x p(x) \log \frac{1}{|X|} = -H(p) - \frac{1}{|X|} \sum_x p(x)$$

- By Pythagorean Property, $D(p,u) = D(p,p^*) + D(p^*,u)$
- KL divergence is nonnegative: $D(r,s) \geq 0$
- Therefore, $D(p,u) \geq D(p^*,u)$

$$-H(p) - \log \frac{1}{|X|} \geq -H(p^*) - \log \frac{1}{|X|}$$

$$H(p) \leq H(p^*)$$

Maximum Entropy Solution: Uniqueness

Suppose some t has the same (maximum) entropy as p^*

$$H(t) = H(p^*)$$

$$D(t, u) = D(p^*, u)$$

$$D(t, p^*) = 0\text{: from Pythagorean Property}$$

$$t = p^*\text{: Definition of KL}$$

Thus p^* is unique

Maximum Likelihood

- Likelihood given the data

$$L(q) = \sum_x \bar{p}(x) \log q(x)$$

- Follow much the same argument as above to show that p^* has the maximum $L(p^*)$ and is unique

$$\text{If } p^* \in P \cap Q, \text{ then } p^* = \operatorname*{argmax}_{q \in Q} L(q)$$

- That is, the same p^* achieves both maximum likelihood (fits data well) and maximum entropy (makes least assumptions)

Generalized Iterative Scaling: 1

Darroch and Ratcliffe 1972

- Requires that the features are nonnegative and sum to a constant, C

$$C = \max_x \sum_j f_j(x)$$

- Add a (k+1)st correction feature, $f_{k+1} \in [0, C]$, such that

$$\forall x\colon f_{k+1} = C - \sum_j f_j(x)$$

- Requires that each event have at least one active feature

$$\forall x\colon \exists j\colon f_j = 1$$

Generalized Iterative Scaling: 2

Iteration counts as superscripts in parentheses

The following iterative procedure converges to $p^* \in P \cap Q$

$$
\begin{aligned}
\alpha_j^{(0)} &= 1 \\
\alpha_j^{(n+1)} &= \alpha_j^{(n)} \left(\frac{E_{\tilde{p}} f_j}{E^{(n)} f_j} \right)^{\frac{1}{C}} \\
E^{(n)} f_j &= \sum_x p^{(n)}(x) f_j(x) \\
p^{(n)}(x) &= \pi \prod_{j=1}^{k+1} \left(\alpha^{(n)} \right)^{f_j(x)}
\end{aligned}
$$

Computing the Iteration

- $E_{\bar{p}} f_j$ is based on the training data (normalized counts of f_j when f_j is binary valued)
- $E^{(n)} f_j$ is intractable if we consider all possible combinations of the features
- Instead, sum over the "contexts" present in the training sample
  - When the classifier seeks to learn the probability of obtaining $a \in A$ given $b \in B$

$$E^{(n)} f_j \approx \sum_{i=1}^{N} \bar{p}(b_i) \sum_{a \in A} p^{(n)}(a|b_i) f_j(a, b_i)$$

- Complexity per iteration: $O(NPA)$, based largely on estimating $E^{(n)} f_j$
  - N: Number of training samples
  - P: Number of predictions
  - A: Average number of features that are active