Density Estimation
Maxent Models


## Entropy

**Definition:** the entropy of a random variable $X$ with probability distribution $p(x) = \Pr[X = x]$ is

$$H(X) = -\mathbb{E}[\log p(X)] = -\sum_x p(x) \log p(x).$$

**Properties:**

- Measure of uncertainty of $p(x)$.
- $H(X) \geq 0$.
- Maximal for uniform distribution. For a finite support, by Jensen’s inequality:

$$H(X) = \mathbb{E}[\log 1/p(X)] \leq \log \mathbb{E}[1/p(X)] = \log N.$$
Relative Entropy

**Definition:** the relative entropy (or Kullback-Leibler divergence) of two distributions $p$ and $q$ is

$$D(p \parallel q) = E_p[\log \frac{p(X)}{q(X)}] = \sum_x p(x) \log \frac{p(x)}{q(x)},$$

with $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$.

**Properties:**

- Assymetric measure of deviation between two distributions. It is convex in $p$ and $q$.
- $D(p \parallel q) \geq 0$ for all $p$ and $q$.
- $D(p \parallel q) = 0$ iff $p = q$. 
Jensen’s Inequality

**Theorem:** Let $X$ be a random variable and $f$ a measurable convex function. Then,

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

**Proof:**

- For a distribution over a finite set, the property follows directly the definition of convexity.
- The general case is a consequence of the continuity of convex functions and the density of finite distributions.
Applications

Non-negativity of relative entropy:

\[-D(p \parallel q) = E_p[\log \frac{q(X)}{p(X)}]\]
\[\leq \log E_p[\frac{q(X)}{p(X)}]\]
\[= \log \sum_x q(x) = 0.\]
Hoeffding’s Bounds

**Theorem:** Let $X_1, X_2, \ldots, X_m$ be a sequence of independent Bernoulli trials taking values in $[0, 1]$, then for all $\epsilon > 0$, the following inequalities hold for $X_m = \frac{1}{m} \sum_{i=1}^{m} X_i$

\[
\Pr[X_m - \mathbb{E}[X_m] \geq \epsilon] \leq e^{-2m\epsilon^2}.
\]

\[
\Pr[X_m - \mathbb{E}[X_m] \leq -\epsilon] \leq e^{-2m\epsilon^2}.
\]
Problem

- **Data**: sample drawn i.i.d. from set $X$ according to some distribution $D$,

  $x_1, \ldots, x_m \in X$.

- **Problem**: find distribution $p$ out of a set $\mathcal{P}$ that best estimates $D$. 
Maximum Likelihood

- **Likelihood**: probability of observing sample under distribution \( p \in \mathcal{P} \), which, given the independence assumption is

\[
\Pr[x_1, \ldots, x_m] = \prod_{i=1}^{m} p(x_i).
\]

- **Principle**: select distribution maximizing sample probability

\[
p_\ast = \arg\max_{p \in \mathcal{P}} \prod_{i=1}^{m} p(x_i),
\]

or

\[
p_\ast = \arg\max_{p \in \mathcal{P}} \sum_{i=1}^{m} \log p(x_i).
\]
Relative Entropy Formulation

- **Empirical distribution**: distribution $\hat{p}$ that assigns to each point the frequency of its occurrence in the sample.

- **Lemma**: $p_\star$ has maximum likelihood $l(p_\star)$ iff

\[
p_\star = \arg\min_{p \in \mathcal{P}} D(\hat{p} \| p).
\]

- **Proof**:

\[
D(\hat{p} \| p) = \sum_{z_i \text{ observed}} \hat{p}(z_i) \log \hat{p}(z_i) - \sum_{z_i \text{ observed}} \hat{p}(z_i) \log p(z_i)
\]

\[
= -H(\hat{p}) - \sum_{z_i \text{ observed}} \frac{\text{count}(z_i)}{N_0} \log p(z_i)
\]

\[
= -H(\hat{p}) - \frac{1}{N_0} \log \prod_{z_i} p(z_i)^{\text{count}(z_i)}
\]

\[
= -H(\hat{p}) - \frac{1}{N_0} l(p).
\]
Maximum A Posteriori (MAP)

- **Principle**: select the most likely hypothesis \( h \in H \) given the sample, with some *prior distribution* over the hypotheses, \( \Pr[h] \),

\[
h_* = \arg\max_{h \in H} \Pr[h | S]
= \arg\max_{h \in H} \frac{\Pr[S | h] \Pr[h]}{\Pr[S]}
= \arg\max_{h \in H} \Pr[S | h] \Pr[h].
\]

- **Note**: for a uniform prior, MAP coincides with maximum likelihood.
Problem

- **Data**: sample drawn i.i.d. from set $X$ according to some distribution $D$,
  $$x_1, \ldots, x_m \in X.$$  

- **Features**: mappings associated to elements of $X$,
  $$\phi_1, \ldots, \phi_N : X \to \mathbb{R}.$$  

- **Problem**: how do we estimate distribution $D$? Uniform distribution $\mu$ over $X$?
Features

Examples: statistical language modeling.
- \( n \)-grams, distance-\( d \) \( n \)-grams.
- class-based \( n \)-grams, word triggers.
- sentence length.
- number and type of verbs.
- various grammatical information (e.g., agreement, POS tags).
- dialog-level information.
Maximum Entropy Principle


- For large $m$, we can give a fairly good estimate of the expected value of each feature (Hoeffding’s inequality):

$$\forall j \in [1, N], \quad \mathbb{E}_{x \sim D} [\phi_j(x)] \approx \frac{1}{m} \sum_{i=1}^{m} \phi_j(x_i).$$

- Find distribution that is closest to the uniform distribution $\mathcal{U}$ and that preserves the expected values of features.

- Closeness is measured using relative entropy (or Kullback-Liebler divergence).
Maximum Entropy Formulation

- **Distributions**: let $\mathcal{P}$ denote the set of distributions

$$\mathcal{P} = \left\{ p : \mathbb{E}_{x \sim p} [\phi_j(x)] = \frac{1}{m} \sum_{i=1}^{m} \phi_j(x_i), j \in [1, n] \right\}.$$  

- **Optimization problem**: find distribution $p_\star$ verifying

$$p_\star = \arg\min_{p \in \mathcal{P}} D(p \| u).$$
Relation with Entropy Maximization

- **Relationship with entropy:**

\[
D(p \parallel u) = \sum_{x \in X} p(x) \log \frac{p(x)}{1/|X|} = \log |X| + \sum_{x \in X} p(x) \log p(x) = \log |X| - H(p).
\]

- **Optimization problem:** convex problem.

\[
\begin{align*}
\text{minimize} & \quad \sum_{x \in X} p(x) \log p(x) \\
\text{subject to} & \quad \sum_{x \in X} p(x) = 1 \\
& \quad \sum_{x \in X} p(x)\phi_j(x) = \frac{1}{m} \sum_{i=1}^{m} \phi_j(x_i), \forall j \in [1, N].
\end{align*}
\]
Maximum Likelihood Gibbs Distrib.

- **Gibbs distributions:** set \( Q \) of distributions defined by

\[
p(x) = \frac{1}{Z} \exp(w \cdot \Phi(x)) = \frac{1}{Z} \exp \left( \sum_{j=1}^{N} w_j \cdot \Phi_j(x) \right),
\]

with \( Z = \sum_{x} \exp(w \cdot \Phi(x)) \).

- **Maximum likelihood with Gibbs distributions:**

\[
p_\ast = \operatorname{argmax}_{q \in \bar{Q}} \sum_{i=1}^{m} \log q(x_i) = \operatorname{argmin}_{q \in \bar{Q}} D(\hat{p} \parallel q),
\]

where \( \bar{Q} \) is the closure of \( Q \).
Duality Theorem

(Della Pietra et al., 1997)

Theorem: Assume that $D(\hat{p} \parallel u) < \infty$. Then, there exists a unique probability distribution $p_\star$ satisfying

1. $p_\star \in \mathcal{P} \cap \bar{\mathcal{Q}}$;
2. $D(p \parallel q) = D(p \parallel p_\star) + D(p_\star \parallel q)$ for any $p \in \mathcal{P}$ and $q \in \bar{\mathcal{Q}}$ (Pythagorean equality);
3. $p_\star = \arg\min_{q \in \bar{\mathcal{Q}}} D(\hat{p} \parallel q)$ (maximum likelihood);
4. $p_\star = \arg\min_{p \in \mathcal{P}} D(p \parallel u)$

Each of these properties determines $p_\star$ uniquely.
Regularization

- **Overfitting:**
  - Features with low counts;
  - Very large feature weights.

- **Approximation parameters:** $\beta_j \geq 0, j \in [1, n]$;

\[ \forall j \in [1, n], \left| \mathbb{E}_{x \sim D} \left[ \phi_j (x) \right] - \frac{1}{m} \sum_{i=1}^{m} \phi_j (x_i) \right| \leq \beta_j. \]
Regularized Maxent

**Optimization problem:**

\[
\begin{align*}
\arg\min_{\mathbf{w}} & \quad \lambda \|\mathbf{w}\|^2 - \frac{1}{m} \sum_{i=1}^{m} \log p_{\mathbf{w}}[x_i], \\
\text{with} & \quad p_{\mathbf{w}}[x] = \frac{1}{Z(x)} \exp(\mathbf{w} \cdot \Phi(x)).
\end{align*}
\]

**Regularization:** $L_1$ or $L_2$ norm, also interpreted as Gaussian or Laplacian priors (Bayesian view). E.g.,

\[
\begin{align*}
\tilde{p}(x) & \leftarrow p_{\mathbf{w}}(x) \Pr[\mathbf{w}] \\
\tilde{l}(\mathbf{w}) & \leftarrow l(\mathbf{w}) - \sum_{j=1}^{n} \frac{w_j^2}{2\sigma_j^2} - \frac{1}{2} \sum_{j=1}^{n} \log(2\pi\sigma_j^2), \\
\text{where} & \quad \Pr[\mathbf{w}] = \prod_{j=1}^{n} \frac{e^{-w_j^2/(2\sigma_j^2)}}{\sqrt{2\pi\sigma_j^2}}.
\end{align*}
\]
Extensions - Bregman Divergences

**Definition:** Let $F$ be a convex and differentiable function, then the Bregman divergence based on $F$ is defined as

$$B_F(y, x) = F(y) - F(x) - (y - x) \cdot \nabla_x F(x).$$

**Examples:**

- Unnormalized relative entropy.
- Euclidean distance.
Conditional Maxent Models

Generalization of Maxent models:

- multi-class classification.
- conditional probability modeling each class.
- different features for each class.
- logistic regression: special case of two classes.
- also known as multinomial logistic regression.
Problem

- **Data:** sample drawn i.i.d. according to some distribution $D$.

  $$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times \{1, \ldots, k\})^m.$$  

- **Multi-label case:**

  $$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times \{-1, +1\}^k)^m.$$  

- **Features:** for any $y \in \{1, \ldots, k\}$, feature vector

  $$x \mapsto \Phi(x, y).$$

- **Problem:** find distribution $\rho$ out of set $\mathcal{P}$ that best estimates the probability distributions $\Pr[y \mid x]$, for all classes $y \in \{1, \ldots, k\}$. 
Conditional Maxent

Optimization problem: maximizing conditional entropy.

\[
\begin{align*}
\text{minimize} & \quad \sum_{y \in [1,k]} \sum_{x \in X} p(y|x) \log p(y|x) \\
\text{subject to} & \quad \forall y \in [1,k], \sum_{x \in X} p(y|x) = 1 \\
& \quad \forall j \in [1,N], \sum_{x \in X} \sum_{y \in [1,k]} p(y|x) \Phi_j(x,y) = \frac{1}{m} \sum_{i=1}^{m} \Phi_j(x_i, y_i).
\end{align*}
\]
Equivalent ML Formulation

- **Optimization problem**: maximize conditional log-likelihood.

\[
\max_{w \in H} \frac{1}{m} \sum_{i=1}^{m} \log p_w[y_i|x_i] \\
\text{subject to: } \forall (x, y) \in X \times [1, k], \\
p_w[y|x] = \frac{1}{Z(x)} \exp(w \cdot \Phi(x, y)) \\
Z(x) = \sum_{y \in Y} \exp(w \cdot \Phi(x, y)).
\]
Regularized Conditional Maxent

**Optimization problem:** maximizing conditional entropy, regularization parameter $\lambda \geq 0$.

$$\min_{w \in H} \lambda \|w\|^2 - \frac{1}{m} \sum_{i=1}^{m} \log p_w[y_i|x_i]$$

subject to: $\forall (x, y) \in X \times [1, k]$,

$$p_w[y|x] = \frac{1}{Z(x)} \exp(w \cdot \Phi(x, y))$$

$$Z(x) = \sum_{y \in Y} \exp(w \cdot \Phi(x, y)).$$

Different norms used for regularization.
Logistic Regression

(Berkson, 1944)

- **Logistic model**: \( k = 2 \).

\[
\Pr[y \mid x] = \frac{1}{Z(x)} e^{w \cdot \Phi(x, y)},
\]

with \( Z(x) = \sum_{y \in [1, k]} e^{w \cdot \Phi(x, y)} \).

- **Properties**:
  
  - linear log-odds ratio (or logit):
    \[
    \log \frac{\Pr[y_2 \mid x]}{\Pr[y_1 \mid x]} = w \cdot (\Phi(x, y_2) - \Phi(x, y_1)).
    \]
  
  - decision rule: sign of log-odds ratio.
    \[
    \Pr[+1 \mid x] = \frac{1}{1 + e^{-w \cdot (\Phi(x, +1) - \Phi(x, -1))}}.
    \]
References


References


