Maximum Entropy Models, Logistic Regression

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu
Motivation

- Probabilistic models:
  - density estimation.
  - classification.
This Lecture

- Notions of information theory.
- Introduction to density estimation.
- Maxent models.
- Conditional Maxent models.
Entropy

(Shannon, 1948)

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

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

Properties:

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

$$H(X) = E\left[\log \frac{1}{p(X)}\right] \leq \log E\left[\frac{1}{p(X)}\right] = \log N.$$
Entropy

- Base of logarithm: not critical; for base 2, \(- \log_2(p(x))\) is the number of bits needed to represent \(p(x)\).

- Definition and notation: the entropy of a distribution \(p\) is defined by the same quantity and denoted by \(H(p)\).

- Special case of Rényi entropy (Rényi, 1961).

- Binary entropy: \(H(p) = -p \log p - (1 - p) \log(1 - p)\).
Relative Entropy

(Shannon, 1948; Kullback and Leibler, 1951)

- **Definition:** the relative entropy (or Kullback-Leibler divergence) between two distributions \( p \) and \( q \) (discrete case) is

\[
D(p \parallel q) = \mathbb{E}_p \left[ \log \frac{p(X)}{q(X)} \right] = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)},
\]

with \( 0 \log \frac{0}{q} = 0 \) and \( p \log \frac{p}{0} = +\infty \).

- **Properties:**
  - asymmetric: in general, \( D(p \parallel q) \neq D(q \parallel p) \) for \( p \neq q \).
  - non-negative: \( D(p \parallel q) \geq 0 \) for all \( p \) and \( q \).
  - definite: \( (D(p \parallel q) = 0) \Rightarrow (p = q) \).
Non-Negativity of Rel. Entropy

By the concavity of log and Jensen's inequality,

\[-D(p \parallel q) = \sum_{x: p(x)>0} p(x) \log \left( \frac{q(x)}{p(x)} \right) \leq \log \left( \sum_{x: p(x)>0} p(x) \frac{q(x)}{p(x)} \right) \]

\[= \log \left( \sum_{x: p(x)>0} q(x) \right) \leq \log(1) = 0.\]
**Bregman Divergence**

(Bregman, 1967)

**Definition:** let $F$ be a convex and differentiable function defined over a convex set $C$ in a Hilbert space $\mathbb{H}$. Then, the Bregman divergence $B_F$ associated to $F$ is defined by

$$B_F(x \parallel y) = F(x) - F(y) - \langle \nabla F(y), x - y \rangle.$$
Bregman Divergence

Examples:

<table>
<thead>
<tr>
<th></th>
<th>$B_F(x \parallel y)$</th>
<th>$F(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Squared $L_2$-distance</td>
<td>$|x - y|^2$</td>
<td>$|x|^2$</td>
</tr>
<tr>
<td>Mahalanobis distance</td>
<td>$(x - y)^\top K^{-1} (x - y)$</td>
<td>$x ^\top K^{-1} x$</td>
</tr>
<tr>
<td>Unnormalized relative entropy</td>
<td>$\widetilde{D}(x \parallel y)$</td>
<td>$\sum_{i \in I} x_i \log x_i - x_i$</td>
</tr>
</tbody>
</table>

- note: relative entropy not a Bregman divergence since not defined over an open set; but, on the simplex, coincides with unnormalized relative entropy

$$\widetilde{D}(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \left[ \frac{p(x)}{q(x)} \right] + (q(x) - p(x)) .$$
Conditional Relative Entropy

Definition: let $p$ and $q$ be two probability distributions over $\mathcal{X} \times \mathcal{Y}$. Then, the conditional relative entropy of $p$ and $q$ with respect to distribution $r$ over $\mathcal{X}$ is defined by

$$\mathbb{E}_{X \sim r} \left[ D(p(\cdot|X) \| q(\cdot|X)) \right] = \sum_{x \in \mathcal{X}} r(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

$$= D(\tilde{p} \| \tilde{q}),$$

with $\tilde{p}(x, y) = r(x)p(y|x)$, $\tilde{q}(x, y) = r(x)q(y|x)$, and the conventions $0 \log 0 = 0$, $0 \log \frac{0}{0} = 0$, and $p \log \frac{p}{0} = +\infty$.

- note: the definition of conditional relative entropy is not intrinsic, it depends on a third distribution $r$. 
This Lecture

- Notions of information theory.
- Introduction to density estimation.
- Maxent models.
- Conditional Maxent models.
Density Estimation Problem

- **Training data:** sample $S$ of size $m$ drawn i.i.d. from set $\mathcal{X}$ according to some distribution $\mathcal{D}$,

  $$S = (x_1, \ldots, x_m).$$

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

- Maximum Likelihood principle: select distribution \( p \in \mathcal{P} \) maximizing likelihood of observed sample \( S \),

\[
p_{\text{ML}} = \arg\max_{p \in \mathcal{P}} \Pr[S \mid p] \\
= \arg\max_{p \in \mathcal{P}} \prod_{i=1}^{m} p(x_i) \\
= \arg\max_{p \in \mathcal{P}} \sum_{i=1}^{m} \log p(x_i).
\]
Relative Entropy Formulation

- **Lemma:** let $\hat{p}_S$ be the empirical distribution for sample $S$, then

$$p_{ML} = \arg\min_{p \in \mathcal{P}} D(\hat{p}_S \parallel p).$$

- **Proof:**

$$D(\hat{p}_S \parallel p) = \sum_x \hat{p}_S(x) \log \hat{p}_S(x) - \sum_x \hat{p}_S(x) \log p(x)$$

$$= -H(\hat{p}_S) - \sum_x \frac{\sum_{i=1}^m 1_{x=x_i}}{m} \log p(x)$$

$$= -H(\hat{p}_S) - \sum_{i=1}^m \sum_x \frac{1_{x=x_i}}{m} \log p(x)$$

$$= -H(\hat{p}_S) - \sum_{i=1}^m \frac{\log p(x_i)}{m}.$$
Maximum a Posteriori (MAP)

- **Maximum a Posteriori principle**: select distribution \( p \in \mathcal{P} \) that is the most likely, given the observed sample \( S \) and assuming a prior distribution \( \Pr[p] \) over \( \mathcal{P} \),

\[
p_{\text{MAP}} = \arg\max_{p \in \mathcal{P}} \Pr[p|S] \\
= \arg\max_{p \in \mathcal{P}} \frac{\Pr[S|p] \Pr[p]}{\Pr[S]} \\
= \arg\max_{p \in \mathcal{P}} \Pr[S|p] \Pr[p].
\]

- **note**: for a uniform prior, ML = MAP.
This Lecture

- Notions of information theory.
- Introduction to density estimation.
- Maxent models.
- Conditional Maxent models.
Density Estimation + Features

- **Training data**: sample $S$ of size $m$ drawn i.i.d. from set $\mathcal{X}$ according to some distribution $\mathcal{D}$,

  $$S = (x_1, \ldots, x_m).$$

- **Features**: associated to elements of $\mathcal{X}$,

  $$\Phi : \mathcal{X} \rightarrow \mathbb{R}^N$$

  $$x \mapsto \Phi(x) = \begin{bmatrix} \Phi_1(x) \\
  \vdots \\
  \Phi_N(x) \end{bmatrix}.$$

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

  - for simplicity, in what follows, $\mathcal{X}$ is assumed to be finite.
Features

Feature functions $\Phi_j$ assumed to be in $H$ and $\|\Phi\|_\infty \leq \Lambda$.

Examples of $H$:

- family of threshold functions $\{x \mapsto 1_{x_i \leq \theta} : x \in \mathbb{R}^N, \theta \in \mathbb{R}\}$ defined over $N$ variables.
- functions defined via decision trees with larger depths.
- $k$-degree monomials of the original features.
- zero-one features (often used in NLP, e.g., presence/absence of a word or POS tag).
Maximum Entropy Principle


- Idea: empirical feature vector average close to expectation. For any $\delta > 0$, with probability at least $1 - \delta$

\[
\left\| \mathbb{E}_{x \sim \mathcal{D}}[\Phi(x)] - \mathbb{E}_{x \sim \tilde{\mathcal{D}}}[\Phi(x)] \right\|_\infty \leq 2\mathcal{R}_m(H) + \Lambda\sqrt{\frac{\log \frac{2}{\delta}}{2m}},
\]

- Maxent principle: find distribution $p$ that is closest to a prior distribution $p_0$ (typically uniform distribution) while verifying

\[
\left\| \mathbb{E}_{x \sim p}[\Phi(x)] - \mathbb{E}_{x \sim \tilde{\mathcal{D}}}[\Phi(x)] \right\|_\infty \leq \beta.
\]

- Closeness is measured using relative entropy.
  - note: no set $\mathcal{P}$ needed to be specified.
Maxent Formulation

Optimization problem:

$$\min_{p \in \Delta} D(p \parallel p_0)$$

subject to: $\left\| E_{x \sim p} [\Phi(x)] - E_{x \sim S} [\Phi(x)] \right\|_{\infty} \leq \beta$.

- convex optimization problem, unique solution.
- $\beta = 0$: standard Maxent (or unregularized Maxent).
- $\beta > 0$: regularized Maxent.
Relation with Entropy

**Relationship with entropy**: for a uniform prior $p_0$,

$$D(p \parallel p_0) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{p_0(x)}$$

$$= -\sum_{x \in \mathcal{X}} p(x) \log p_0(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x)$$

$$= \log |\mathcal{X}| - H(p).$$
Maxent Problem

Optimization: convex optimization problem.

\[
\min_p \sum_{x \in \mathcal{X}} p(x) \log p(x)
\]

subject to: 
\[
p(x) \geq 0, \forall x \in \mathcal{X}
\]
\[
\sum_{x \in \mathcal{X}} p(x) = 1
\]
\[
\left| \sum_{x \in \mathcal{X}} p(x) \Phi_j(x) - \frac{1}{m} \sum_{i=1}^{m} \Phi_j(x_i) \right| \leq \beta, \forall j \in [1, N].
\]
Gibbs Distributions

Gibbs distributions: set $Q$ of distributions $p_w$ with $w \in \mathbb{R}^N$,

$$p_w[x] = \frac{p_0[x] \exp (w \cdot \Phi(x))}{Z} = \frac{p_0[x] \exp (\sum_{j=1}^{N} w_j \Phi_j(x))}{Z},$$

with $Z = \sum_x p_0[x] \exp (w \cdot \Phi(x))$.

Rich family:

- for linear and quadratic features: includes Gaussians and other distributions with non-PSD quadratic forms in exponents.
- for higher-degree polynomials of raw features: more complex multi-modal distributions.
Examples

\begin{align*}
    p[(x_1, x_2)] &= e^{-\left(x_1^2 + x_2^2\right)} / Z. \\
    p[(x_1, x_2)] &= e^{-\left(x_1^4 + x_2^4\right) + x_1^2 - x_2^2} / Z.
\end{align*}
Dual Problems

- Regularized Maxent problem:

\[
\min_p F(p) = \overline{D}(p \parallel p_0) + I_C(\mathbb{E}[\Phi]),
\]

\[
\overline{D}(p \parallel p_0) = D(p \parallel p_0) \text{ if } p \in \Delta, +\infty \text{ otherwise};
\]

with \[
C = \left\{ u : \| u - \mathbb{E}_S[\Phi] \|_\infty \leq \beta \right\};
\]

\[
I_C(x) = 0 \text{ if } x \in C, I_C(x) = +\infty \text{ otherwise}.
\]

- Regularized Maximum Likelihood problem with Gibbs distributions:

\[
\sup_w G(w) = \frac{1}{m} \sum_{i=1}^m \log \left[ \frac{p_w[x_i]}{p_0[x_i]} \right] - \beta \| w \|_1.
\]
Duality Theorem

(Della Pietra et al., 1997; Dudík et al., 2007; Cortes et al., 2015)

- **Theorem**: the regularized Maxent and ML with Gibbs distributions problems are equivalent,

\[
\sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w}) = \min_{\mathbf{p}} F(\mathbf{p}).
\]

- Furthermore, let \( \mathbf{p}^* = \arg \min_{\mathbf{p}} F(\mathbf{p}) \), then, for any \( \epsilon > 0 \),

\[
\left( |G(\mathbf{w}) - \sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w})| < \epsilon \right) \Rightarrow \left( D(\mathbf{p}^* \parallel \mathbf{p}_\mathbf{w}) \leq \epsilon \right).
\]
Maxent formulation:

- no explicit restriction to a family of distributions $\mathcal{P}$.
- but solution coincides with regularized ML with a specific family $\mathcal{P}$!
- more general Bregman divergence-based formulation.
L₁-Regularized Maxent

(Kazama and Tsujii, 2003)

- Optimization problem:

\[
\inf_{\mathbf{w} \in \mathbb{R}^N} \beta \|\mathbf{w}\|_1 - \frac{1}{m} \sum_{i=1}^{m} \log p_{\mathbf{w}}[x_i].
\]

where \( p_{\mathbf{w}}[x] = \frac{1}{Z} \exp(\mathbf{w} \cdot \Phi(x)) \).

- Bayesian interpretation: equivalent to MAP with Laplacian prior \( q_{\text{prior}}(\mathbf{w}) \) (Williams, 1994),

\[
\max_{\mathbf{w}} \log \left( \prod_{i=1}^{m} p_{\mathbf{w}}[x_i] q_{\text{prior}}(\mathbf{w}) \right)
\]

with \( q_{\text{prior}}(\mathbf{w}) = \prod_{j=1}^{N} \frac{\beta_j}{2} \exp(-\beta_j |w_j|) \).
Generalization Guarantee

(Dudík et al., 2007)

- **Notation:** $\mathcal{L}_D(w) = \mathbb{E}_{x \sim D}[-\log p_w(x)]$, $\mathcal{L}_S(w) = \mathbb{E}_{x \sim S}[-\log p_w(x)]$.

- **Theorem:** Fix $\delta > 0$. Let $\hat{w}$ be the solution of the L1-reg. Maxent problem for $\beta = 2\mathfrak{R}_m(H) + \Lambda \sqrt{\log(\frac{2}{\delta})/2m}$. Then, with probability at least $1 - \delta$,

$$
\mathcal{L}_D(\hat{w}) \leq \inf_w \mathcal{L}_D(w) + 2\|w\|_1 \left[2\mathfrak{R}_m(H) + \Lambda \sqrt{\frac{\log \frac{2}{\delta}}{2m}}\right].
$$
Proof

By Hölder’s inequality and the concentration bound for average feature vectors,

\[ \mathcal{L}_D(\hat{w}) - \mathcal{L}_S(\hat{w}) = \hat{w} \cdot [E[\Phi] - \bar{E}[\Phi]] \]

\[ \leq \| \hat{w} \|_1 \| E[\Phi] - \bar{E}[\Phi] \|_\infty \leq \beta \| \hat{w} \|_1. \]

Since \( \hat{w} \) is a minimizer,

\[ \mathcal{L}_D(\hat{w}) - \mathcal{L}_D(w) = \mathcal{L}_D(\hat{w}) - \mathcal{L}_S(\hat{w}) + \mathcal{L}_S(\hat{w}) - \mathcal{L}_D(w) \]

\[ \leq \beta \| \hat{w} \|_1 + \mathcal{L}_S(\hat{w}) - \mathcal{L}_D(w) \]

\[ \leq \beta \| w \|_1 + \mathcal{L}_S(w) - \mathcal{L}_D(w) \leq 2\beta \| w \|_1. \]
**L₂-Regularized Maxent**

(Chen and Rosenfeld, 2000; Lebanon and Lafferty, 2001)

- **Different relaxations:**
  - **L₁ constraints:**
    \[
    \forall j \in [1, N], \quad \left| \mathbb{E}_{x \sim p}[\Phi_j(x)] - \mathbb{E}_{x \sim \hat{p}}[\Phi_j(x)] \right| \leq \beta_j.
    \]
  - **L₂ constraints:**
    \[
    \left\| \mathbb{E}_{x \sim p}[\Phi(x)] - \mathbb{E}_{x \sim \hat{p}}[\Phi(x)] \right\|_2 \leq B.
    \]
L₂-Regularized Maxent

Optimization problem:

\[ \inf_{\mathbf{w} \in \mathbb{R}^N} \beta \| \mathbf{w} \|_2^2 - \frac{1}{m} \sum_{i=1}^{m} \log p_{\mathbf{w}}[x_i]. \]

where \( p_{\mathbf{w}}[x] = \frac{1}{Z} \exp (\mathbf{w} \cdot \Phi(x)) \).

Bayesian interpretation: equivalent to MAP with Gaussian prior \( q_{\text{prior}}(\mathbf{w}) \) (Goodman, 2004),

\[ \max_{\mathbf{w}} \log \left( \prod_{i=1}^{m} p_{\mathbf{w}}[x_i] q_{\text{prior}}(\mathbf{w}) \right) \]

with \( q_{\text{prior}}(\mathbf{w}) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{w_j^2}{2\sigma^2}}. \)
This Lecture

- Notions of information theory.
- Introduction to density estimation.
- Maxent models.
- Conditional Maxent models.
Conditional Maxent Models

Maxent models for conditional probabilities:

- conditional probability modeling each class.
- use in multi-class classification.
- can use different features for each class.
- a.k.a. multinomial logistic regression.
- logistic regression: special case of two classes.
Problem

- **Data**: sample drawn i.i.d. according to some distribution $D$, 
  \[ S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m. \]

- $\mathcal{Y} = \{1, \ldots, k\}$, or $\mathcal{Y} = \{0, 1\}^k$ in multi-label case.

- **Features**: mapping $\Phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^N$.

- **Problem**: find accurate conditional probability models $\Pr[\cdot | x], x \in \mathcal{X}$, based on $\Phi$. 
Conditional Maxent Principle

(Berger et al., 1996; Cortes et al., 2015)

- **Idea**: empirical feature vector average close to expectation. For any $\delta > 0$, with probability at least $1 - \delta$,

$$\left\| \mathbb{E}_{x \sim \hat{p}} [\Phi(x, y)] - \mathbb{E}_{x \sim \hat{p}} [\Phi(x, y)] \right\|_\infty \leq 2\mathcal{R}_m(H) + \sqrt{\frac{\log 2}{2m}}.$$  

- **Maxent principle**: find conditional distributions $p[\cdot| x]$ that are closest to priors $p_0[\cdot| x]$ (typically uniform distributions) while verifying

$$\left\| \mathbb{E}_{x \sim \hat{p}} [\Phi(x, y)] - \mathbb{E}_{x \sim \hat{p}} [\Phi(x, y)] \right\|_\infty \leq \beta.$$  

- Closeness is measured using conditional relative entropy based on $\hat{p}$.  

Cond. Maxent Formulation

(Berger et al., 1996; Cortes et al., 2015)

- **Optimization problem:** find distribution \( p \) solution of

\[
\min_{p[\cdot|x] \in \Delta} \sum_{x \in \mathcal{X}} \hat{p}[x] D(p[\cdot|x] \| p_0[\cdot|x])
\]

\[
\text{s.t. } \left\| \mathbb{E}_{x \sim \hat{p}} \left[ \mathbb{E}_{y \sim p[\cdot|x]} [\Phi(x, y)] \right] - \mathbb{E}_{(x,y) \sim S} [\Phi(x, y)] \right\|_\infty \leq \beta.
\]

- convex optimization problem, unique solution.
- \( \beta = 0 \): unregularized conditional Maxent.
- \( \beta > 0 \): regularized conditional Maxent.
**Dual Problems**

- Regularized conditional Maxent problem:
  \[
  \tilde{F}(p) = \mathbb{E}_{x \sim \hat{p}} \left[ \tilde{D}(p[\cdot|x] \parallel p_0[\cdot|x]) + I_\Delta(p[\cdot|x]) \right] + I_C \left( \mathbb{E}_{x \sim \hat{p}} [\Phi] \right).
  \]

- Regularized Maximum Likelihood problem with conditional Gibbs distributions:
  \[
  \tilde{G}(w) = \frac{1}{m} \sum_{i=1}^{m} \log \frac{p_w[y_i|x_i]}{p_0[y_i|x_i]} - \beta \|w\|_1,
  \]
  where \(\forall (x, y) \in \mathcal{X} \times \mathcal{Y},\)

  \[
  p_w[y|x] = \frac{p_0[y|x] \exp(w \cdot \Phi(x, y))}{Z(x)}
  \]

  \[
  Z(x) = \sum_{y \in \mathcal{Y}} p_0[y|x] \exp(w \cdot \Phi(x, y)).
  \]
Duality Theorem

Theorem: the regularized conditional Maxent and ML with conditional Gibbs distributions problems are equivalent,

\[ \sup_{\mathbf{w} \in \mathbb{R}^N} \tilde{G}(\mathbf{w}) = \min_{\mathbf{p}} \tilde{F}(\mathbf{p}). \]

Furthermore, let \( p^* = \arg\min_{\mathbf{p}} \tilde{F}(\mathbf{p}) \), then, for any \( \epsilon > 0 \),

\[ \left( |\tilde{G}(\mathbf{w}) - \sup_{\mathbf{w} \in \mathbb{R}^N} \tilde{G}(\mathbf{w})| < \epsilon \right) \Rightarrow \mathbb{E}_{\mathbf{x} \sim \hat{p}} \left[ D(p^*[\cdot|\mathbf{x}] \| p_{\mathbf{w}}[\cdot|\mathbf{x}]) \right] \leq \epsilon. \]
Regularized Cond. Maxent

(Berger et al., 1996; Cortes et al., 2015)

- **Optimization problem**: convex optimizations, regularization parameter $\lambda \geq 0$.

\[
\min_{w \in \mathbb{R}^N} \lambda \|w\|_1 - \frac{1}{m} \sum_{i=1}^{m} \log p_w[y_i|x_i]
\]

or

\[
\min_{w \in \mathbb{R}^N} \lambda \|w\|_2^2 - \frac{1}{m} \sum_{i=1}^{m} \log p_w[y_i|x_i],
\]

where $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$,

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

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

Foundations of Machine Learning
More Explicit Forms

**Optimization problem:**

\[
\min_{\mathbf{w} \in \mathbb{R}^N} \left\{ \lambda \left\| \mathbf{w} \right\|_1^2 + \frac{1}{\lambda \left\| \mathbf{w} \right\|_2^2} \sum_{i=1}^m \log \left[ \sum_{y \in \mathcal{Y}} \exp \left( \mathbf{w} \cdot \Phi(x_i, y) - \mathbf{w} \cdot \Phi(x_i, y_i) \right) \right] \right\}.
\]

\[
\min_{\mathbf{w} \in \mathbb{R}^N} \left\{ \lambda \left\| \mathbf{w} \right\|_1^2 - \mathbf{w} \cdot \frac{1}{m} \sum_{i=1}^m \Phi(x_i, y_i) + \frac{1}{m} \sum_{i=1}^m \log \left[ \sum_{y \in \mathcal{Y}} e^{\mathbf{w} \cdot \Phi(x_i, y)} \right] \right\}.
\]
Related Problem

- **Optimization problem**: log-sum-exp replaced by max.

\[
\min_{w \in \mathbb{R}^N} \left\{ \lambda \| w \|_1 + \frac{1}{m} \sum_{i=1}^{m} \max_{y \in \mathcal{Y}} \left( w \cdot \Phi(x_i, y) - w \cdot \Phi(x_i, y_i) \right) \right\}.
\]
Common Feature Choice

**Multi-class features:**

\[
\Phi(x, y) = \begin{bmatrix}
0 \\
\vdots \\
0 \\
0 \\
\vdots \\
0
\end{bmatrix} \Gamma(x) = \begin{bmatrix}
w_1 \\
\vdots \\
w_{y-1} \\
w_y \\
w_{y+1} \\
w_{|\mathcal{Y}|}
\end{bmatrix}
\Rightarrow w \cdot \Phi(x, y) = w_y \cdot \Gamma(x).
\]

**L2-regularized cond. maxent optimization:**

\[
\min_{\mathbf{w} \in \mathbb{R}^N} \lambda \sum_{y \in \mathcal{Y}} \|w_y\|_2^2 + \frac{1}{m} \sum_{i=1}^{m} \log \left[ \sum_{y \in \mathcal{Y}} \exp \left( w_y \cdot \Gamma(x_i) - w_{y_i} \cdot \Gamma(x_i) \right) \right].
\]
Prediction

Prediction with \( p_w[y|x] = \frac{\exp(w \cdot \Phi(x,y))}{Z(x)} \):

\[ \hat{y}(x) = \arg\max_{y \in \mathcal{Y}} p_w[y|x] = \arg\max_{y \in \mathcal{Y}} w \cdot \Phi(x, y). \]
Binary Classification

Simpler expression:

\[
\sum_{y \in \mathcal{Y}} \exp \left( \mathbf{w} \cdot \Phi(x_i, y) - \mathbf{w} \cdot \Phi(x_i, y_i) \right) \\
= e^{\mathbf{w} \cdot \Phi(x_i, +1) - \mathbf{w} \cdot \Phi(x_i, y_i)} + e^{\mathbf{w} \cdot \Phi(x_i, -1) - \mathbf{w} \cdot \Phi(x_i, y_i)} \\
= 1 + e^{-y_i \mathbf{w} \cdot [\Phi(x_i, +1) - \Phi(x_i, -1)]} \\
= 1 + e^{-y_i \mathbf{w} \cdot \Psi(x_i)},
\]

with \( \Psi(x) = \Phi(x, +1) - \Phi(x, -1) \).
Logistic Regression

- Binary case of conditional Maxent.

- Optimization problem:

\[
\min_{\mathbf{w} \in \mathbb{R}^N} \left\{ \lambda \| \mathbf{w} \|_1 + \frac{1}{m} \sum_{i=1}^{m} \log \left[ 1 + e^{-y_i \mathbf{w} \cdot \Psi(x_i)} \right] \right\}.
\]

- convex optimization.

- variety of solutions: SGD, coordinate descent, etc.

- coordinate descent: similar to AdaBoost with logistic loss \( \phi(-u) = \log_2(1 + e^{-u}) \geq 1_{u \leq 0} \) instead of exponential loss.
Theorem: assume that $\pm \Phi_j \in H$ for all $j \in [1, N]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of a sample $S$ of size $m$, for all $f : x \mapsto w \cdot \Phi(x)$,

$$R(f) \leq \frac{1}{m} \sum_{i=1}^{m} \log u_0 \left( 1 + e^{-y_i w \cdot \Phi(x_i)} \right) + 4\|w\|_1 \mathcal{R}_m(H)$$

$$+ \sqrt{\frac{\log \log_2 2\|w\|_1}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{m}},$$

where $u_0 = 1 + \frac{1}{e}$. 

Generalization Bound
Proof

Proof: by the learning bound for convex ensembles holding uniformly for all \( \rho \), with probability at least \( 1 - \delta \), for all \( f \) and \( \rho > 0 \),

\[
R(f) \leq \frac{1}{m} \sum_{i=1}^{m} \frac{y_i \mathbf{w} \cdot \Phi(x_i)}{\rho \|\mathbf{w}\|_1} - 1 \leq 0 + \frac{4}{\rho} \mathcal{R}_m(H) + \sqrt{\frac{\log \log_2 \frac{2}{\delta}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{m}}.
\]

Choosing \( \rho = \frac{1}{\|\mathbf{w}\|_1} \) and using \( 1_{u \leq 1} \leq \log_{u_0} (1 + e^{-u}) \) yields immediately the learning bound of the theorem.
Logistic Regression

Logistic model:

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

where \( Z(x) = e^{\mathbf{w} \cdot \Phi(x, +1)} + e^{\mathbf{w} \cdot \Phi(x, -1)} \)

Properties:

- linear decision rule, sign of log-odds ratio:
  \[
  \log \frac{\Pr[y = +1 \mid x]}{\Pr[y = -1 \mid x]} = \mathbf{w} \cdot (\Phi(x, +1) - \Phi(x, -1)) = \mathbf{w} \cdot \Psi(x).
  \]

- logistic form:
  \[
  \Pr[y = +1 \mid x] = \frac{1}{1 + e^{-\mathbf{w} \cdot [\Phi(x, +1) - \Phi(x, -1)]}} = \frac{1}{1 + e^{-\mathbf{w} \cdot \Psi(x)}}.
  \]
Logistic/Sigmoid Function

\[ f: x \mapsto \frac{1}{1 + e^{-x}} \]

\[ \Pr[y = +1 \mid x] = f(w \cdot \Psi(x)). \]
Applications

- **Natural language processing** (Berger et al., 1996; Rosenfeld, 1996; Pietra et al., 1997; Malouf, 2002; Manning and Klein, 2003; Mann et al., 2009; Ratnaparkhi, 2010).

- **Species habitat modeling** (Phillips et al., 2004, 2006; Dudík et al., 2007; Elith et al., 2011).

- **Computer vision** (Jeon and Manmatha, 2004).
Extensions

- Extensive theoretical study of alternative regularizations: (Dudík et al., 2007) (see also (Altun and Smola, 2006) though some proofs unclear).

- Maxent models with other Bregman divergences (see for example (Altun and Smola, 2006)).

- Structural Maxent models (Cortes et al., 2015):
  - extension to the case of multiple feature families.
  - empirically outperform Maxent and L1-Maxent.
  - conditional structural Maxent: coincide with deep boosting using the logistic loss.
Conclusion

- Logistic regression/maxent models:
  - theoretical foundation.
  - natural solution when probabilities are required.
  - widely used for density estimation/classification.
  - often very effective in practice.
  - distributed optimization solutions.
  - no natural non-linear L1-version (use of kernels).
  - connections with boosting.
  - connections with neural networks.
References


References


References


References

