Maximum Entropy Framework

Sridhar Mahadevan
mahadeva@cs.umass.edu
University of Massachusetts

Topics

- Maximum entropy is the dual of maximum likelihood.
- ME involves finding the “maximally general” distribution that satisfies a set of prespecified constraints.
- The notion of “maximal generality” is made precise using the concept of entropy.
- The distribution that satisfies the ME constraint can be shown to be in an exponential form.
- The concept of maximum entropy comes from statistical physics, and this whole area of machine learning borrows heavily from physics.
- Undirected graphical models are called random field models, and also derive much of their formalism from statistical physics.
The concept of entropy was investigated originally in statistical physics and thermodynamics.

The temperature of a gas is proportional to the average kinetic energy of the molecules in the gas.

The distribution of velocities at a given temperature is a maximum entropy distribution (also known as the Maxwell-Boltzmann distribution).

*Boltzmann machines* are a type of neural network that get their inspiration from statistical physics. There is a whole subfield of energy based models in machine learning based on borrowing ideas from statistical physics.
Maximum Entropy Example

- Consider estimating a joint distribution over two variables $u$ and $v$, given some constraints.

<table>
<thead>
<tr>
<th>$P(u,v)$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>1.0</td>
</tr>
</tbody>
</table>

- The maximum entropy framework suggests picking values that make the least commitments, while being consistent with the constraints that are given.
- For this example, this suggests a unique assignment of values, but in general, determining the set of possible distributions is not so straightforward.

Maximum entropy distributions

- Consider maximizing the entropy $h(P)$ over all distributions $P$ satisfying the following constraints:
  - $P(x) \geq 0$ (where $x \in S$, the support of $P$).
  - $\sum_{x \in S} P(x) = 1$
  - $\sum_{x \in S} P(x)r_i(x) = \alpha_i$ for $1 \leq i \leq m$. 
Maximum entropy distribution

- Writing out the Lagrangian, we get

\[ \Lambda(P) = -\sum_{x \in S} P(x) \ln P(x) + \lambda_0 \left( \sum_{x \in S} P(x) - 1 \right) + \sum_{i=1}^{m} \lambda_i \left( \sum_{x \in S} P(x)r_i(x) - \alpha_i \right) \]

- Taking the gradient of this Lagrangian w.r.t. \( P(x) \) we get

\[ \frac{\partial}{\partial P} \Lambda(P) = \left( -\ln P(x) - 1 + \lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x) \right) \]

\[ \Rightarrow P(x) = e^{\left( \lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x) \right)} \quad x \in S \]

where \( \lambda_0, \lambda_1, \ldots, \lambda_m \) are chosen so that the constraints are satisfied.

---

Differential Entropy

- If \( X \) is a continuous variable, the entropy \( H(X) \) is referred to as the differential entropy.

- A similar result can be proved for continuous variables.

  - \( P(x) \geq 0 \) (where \( P(x) = 0 \) if \( x \) is not in the support \( S \) of \( P \)).
  - \( \int_S P(x) dx = 1 \)
  - \( \int_S P(x)r_i(x) = \alpha_i \) for \( 1 \leq i \leq m \).

- The Lagrangian in the continuous case is

\[ \Lambda(P) = -\int_S P(x) \ln P(x) + \lambda_0 \int_S P(x) dx + \sum_{i=1}^{m} \lambda_i \left( \int_S P(x)r_i(x) dx - \alpha_i \right) \]
Example 1: Uniform Distribution

- Find the maximum entropy distribution that satisfies the following constraints:
  - \( S = [a, b] \)
- Since there are no other constraints, the form of the distribution must be
  \[
P(x) = e^{\lambda_0}
  \]
  This is because all \( \lambda_i = 0, i > 0 \).
- Solving this integral, we find
  \[
  \int_a^b e^{\lambda_0} \, dx = 1 \Rightarrow e^{\lambda_0}(b - a) = 1
  \]
  which immediately gives us the uniform distribution because
  \[
P(x) = e^{\lambda_0} = \frac{1}{b - a}
  \]

Example 2: Gaussian Distribution

- Find the maximum entropy distribution that satisfies the following constraints:
  - \( S = (-\infty, \infty) \)
  - \( EX = 0 \) or \( \int_{-\infty}^{\infty} xP(x) \, dx = 0 \)
  - \( EX^2 = \sigma^2 \).
- The form of our distribution in this case is
  \[
P(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 x^2}
  \]
  We have to find the constants \( \lambda_i \) by solving
  \[
  \int_{-\infty}^{\infty} e^{\lambda_0 + \lambda_1 x + \lambda_2 x^2} \, dx = 1
  \]
  \[
  \int_{-\infty}^{\infty} xe^{\lambda_0 + \lambda_1 x + \lambda_2 x^2} \, dx = 0
  \]
  \[
  \int_{-\infty}^{\infty} x^2 e^{\lambda_0 + \lambda_1 x + \lambda_2 x^2} \, dx = \sigma^2
  \]
- Show by choosing \( \lambda_1 = 0, \lambda_2 = -\frac{1}{2\sigma^2} \) and \( \lambda_0 = -\ln \sqrt{2\pi\sigma} \), we get the Gaussian distribution.
Example 3: Mystery Distribution

- Find the maximum entropy distribution that satisfies the following constraints:
  - \( S = [0, \infty) \)
  - \( EX = \mu \)
- The solution to this mystery distribution gives us a well-known PDF (highly recommended you solve this by finding the \( \lambda_i \) constants!)

Cross Entropy and Log Likelihood

- Let us define the empirical distribution \( \tilde{P}(x) \) by placing a point mass at each data point

\[
\tilde{P}(x) = \frac{1}{N} \sum_{n=1}^{N} I(x, x_n)
\]

- The cross-entropy of the empirical distribution with the model \( P(x|\theta) \) gives:

\[
\sum_{x} \tilde{P}(x) \log P(x|\theta) = \sum_{x} \frac{1}{N} \sum_{n=1}^{N} I(x, x_n) \log P(x|\theta)
\]

\[
= \frac{1}{N} \sum_{n=1}^{N} \sum_{x} I(x, x_n) \log P(x|\theta)
\]

\[
= \frac{1}{N} \sum_{n=1}^{N} \log P(x_n|\theta) = \frac{1}{N} l(\theta|X)
\]
KL Divergence and Maximum Likelihood

Let us compute the KL divergence between the empirical distribution $\tilde{P}(x)$ and the model $P(x|\theta)$.

\[
KL(\tilde{P}||P(x|\theta)) = \sum_x \tilde{P}(x) \log \frac{\tilde{P}(x)}{P(x|\theta)}
= \sum_x \tilde{P}(x) \log \tilde{P}(x) - \sum_x \tilde{P}(x) \log P(x|\theta)
= \sum_x \tilde{P}(x) \log \tilde{P}(x) - \frac{1}{N} l(\theta|X)
\]

- The first term is independent of the model parameters.
- This means optimizing the parameters $\theta$ to minimize the KL divergence between the model and the empirical distribution is the same as maximizing the (log) likelihood.

Conditional Maximum Entropy

In classification, we are interested in modeling the conditional distribution $P(y|x)$. Let us now turn to viewing classification in the maximum entropy framework.

To model $P(y|x)$, let us assume we are given a set of binary features $f_i(x, y)$, which represents the co-occurrence of the ("contextual") input $x$ and (label) output $y$.

For example, in a natural language application, $x$ may be the set of words surrounding a given word, and $y$ may represent the part of speech (or a category like “company name”).

Let us generalize the idea from logistic regression, and construct an exponential model $P(y|x, \lambda)$ using a weighted linear combination of the features $f_i(x, y)$, where the weights are scalars $\lambda_i$.

We want to set up constraints based on the “error” between the empirical joint distribution $\tilde{P}(x, y)$ and the product model distribution $\tilde{P}(x)P(y|x, \lambda)$. 
Constraints

We now define a set of constraints on the simplex of probability distributions.

Each constraint is based on comparing the prediction of the model, and the empirical distribution based on the training set \((x_1, y_1), \ldots, (x_n, y_n)\).

Let us define the empirical joint distribution \(\tilde{P}(u, v)\) as simply the frequency of co-occurrence of \((u, v)\) in the training set.

\[
\tilde{P}(u, v) = \frac{1}{N} \sum_{x, y} I((u, v), (x, y))
\]

For each binary valued feature \(f(x, y)\) we can impose an additional constraint, namely the expected value of \(f\) under the model should agree with its expectation under the empirical joint distribution.

\[
E_P(f) \equiv \sum_{x, y} \tilde{P}(x)P(y|x, \lambda)f(x, y) = E_{\tilde{P}}(f) \equiv \sum_{x, y} \tilde{P}(x, y)f(x, y)
\]

Maximum Conditional Entropy

Given \(n\) feature functions \(f_i(x, y)\), we would like our estimated model to be in accordance with the sample distribution

\[
C = \left( P \in \mathcal{P} | E_P(f_i(x, y)) = E_{\tilde{P}}(f_i(x, y)) \text{ for } i \in \{1, \ldots, n\} \right)
\]

But these set of constraints still define a simplex of possible distributions, and we need to constrain this further.

The maximum conditional entropy principle states that we should pick the model that maximizes the conditional entropy

\[
H(P) = - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \tilde{P}(x)P(y|x, \lambda) \log P(y|x, \lambda)
\]

Putting both these constraints together, we get

\[
P^* = \arg\max_{P \in C} H(P)
\]
Constrained Optimization

For each feature function $f_i(x, y)$, we introduce a Lagrangian multiplier parameter $\lambda_i$, and form the overall Lagrangian as

$$\Lambda(P, \lambda) \equiv H(P) + \sum_i \lambda_i \left( E_P(f_i) - E_{\tilde{P}}(f_i) \right)$$

If we keep $\lambda$ fixed, we can find the unconstrained maximum of the Lagrangian $\Lambda(P, \lambda)$ over all $P \in \mathcal{P}$. Let $P_\lambda$ denote the argument where the maximum is reached, and $\Psi(\lambda)$ denote the maximizing value

$$P_\lambda = \arg\max_{P \in \mathcal{P}} \Lambda(P, \lambda)$$

$$\Psi(\lambda) = \Lambda(P_\lambda, \lambda)$$

Maximum Entropy Distribution

**Key theorem**: By solving the Lagrangian above, one can show that the maximum entropy distribution is given by the following exponential form

$$P(y|x, \lambda) = \frac{1}{Z_\lambda} e^{\left( \sum_i \lambda_i f_i(x, y) \right)}$$

where $Z_\lambda(x) = \sum_y e^{\left( \sum_i \lambda_i f_i(x, y) \right)}$

The maximizing value is

$$\Psi(\lambda) \equiv \Lambda(P_\lambda, \lambda) = -\sum_x \tilde{P}(x) \log Z_\lambda(x) + \sum_i \lambda_i E_{\tilde{P}}(f_i)$$
Karush Kuhn Tucker Theorem

- The original constrained optimization problem, called the *primal* is
  \[ P^* = \arg\max_{P \in C} H(P) \]

- The *dual* (unconstrained) optimization problem is
  \[ \text{Find } \lambda^* = \arg\max_{\lambda} \Psi(\lambda) \]

- The KKT theorem shows that under certain conditions, the solution to the dual problem gives us the solution to the primal problem.
- More precisely, \( P^* = P_{\lambda^*} \)
- We will see the KKT theorem shows up again in support vector machines.

Maximum Likelihood Revisited

- The log-likelihood of the empirical distribution \( \tilde{P} \), as predicted by the discriminant function \( P(Y|X, \lambda) \) is
  \[ L_{\tilde{P}}(P) = \log \prod_{x,y} P(y|x, \lambda) \tilde{P}(x, y) = \sum_{x,y} \tilde{P}(x, y) \log P(y|x, \lambda) \]

- We can show that the log-likelihood above is exactly the dual function \( \Psi(\lambda) \) given above.
- In other words, the maximum entropy model \( P^* \in C \) is exactly the model in the parametric family \( P(Y|X, \lambda) \) that maximizes the likelihood of the training instances \( \tilde{P} \).
Let us see what the maximum likelihood solution yields, by taking the gradient

\[
\frac{\partial L_{\hat{P}}(P)}{\partial \lambda_i} = \frac{\partial}{\partial \lambda_i} \left( \log \prod_{x,y} P(y|x, \lambda) \hat{P}(x,y) \right) = \frac{\partial}{\partial \lambda_i} \left( \sum_{x,y} \hat{P}(x,y) \log P(y|x, \lambda) \right) = \sum_{x,y} \hat{P}(x,y) f_i(x,y) - \sum_{x,y} \hat{P}(x) P(y|x, \lambda) f_i(x,y)
\]

We now see that this is exactly the same set of constraints that the maximum entropy solution was based on.
Maximum Entropy Parameter Estimation

- Note that maximum likelihood of exponential model is not analytically solvable, since the normalizer couples the parameters together

\[
\log \Psi(\lambda) = \sum_i \lambda_i \sum_{x, y} \tilde{P}(x, y) f_i(x, y) - \sum_x \tilde{P}(x) \log \sum_y e^{\sum_i \lambda_i f_i(x, y)}
\]

- Methods based on Generalized Iterative Scaling [Darroch and Ratcliff, ’72; Berger, Della Pietra, Della Pietra]
  - Use Jensen’s inequality and a bound on \(-\log x \geq 1 - x\) to derive a lower bound on log-likelihood improvement
  - Introduce \(f^\#(x, y) = \sum_i f_i(x, y)\) to get a PDF-like term by averaging out \(f(x, y)\).

- Gradient methods
  - Plug in favorite first-order or second-order gradient method from numerical analysis, including Newton’s method, conjugate gradient, line search
  - These have been recently found to be much faster than GIS for logistic regression [Minka 2003]

Generative vs Discriminative Models
Distributions on Undirected Graphs

- Earlier we saw that for directed graphical models, d-separation formalizes the notion of conditional independence.
- We also saw that the joint distribution factorizes as a product of terms, each of which involves a node and its parents. The factorization property of directed graphs ensures that every distribution defined in this way must satisfy all d-separation properties in the graph.
- A set of nodes \( A \) is conditionally independent of a set of nodes \( B \) given the separating set \( C \), if every path from a node in \( A \) to a node in \( B \) goes through a node in \( C \).
- We say that a probability distribution \( P \) is globally Markov with respect to an undirected graph \( G \) iff for every disjoint set of nodes \( A, B, \) and \( C \), if \( A \perp B \mid C \), then the distribution also satisfies the same property.

Two Node Chain Graph

\[
P(A, B, C) = \frac{1}{Z} \phi_1(A, B) \cdot \phi_2(B, C)
\]

\[
Z = \sum_{A,B,C} \phi_1(A, B) \cdot \phi_2(B, C)
\]
Two Node Chain Graph

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$A^0$</td>
<td>$B^0$</td>
<td>$C^0$</td>
<td>10</td>
<td>0.0413</td>
</tr>
<tr>
<td>$A^0$</td>
<td>$B^0$</td>
<td>$C^1$</td>
<td>100</td>
<td>0.413</td>
</tr>
<tr>
<td>$A^0$</td>
<td>$B^1$</td>
<td>$C^0$</td>
<td>1</td>
<td>0.00413</td>
</tr>
<tr>
<td>$A^0$</td>
<td>$B^1$</td>
<td>$C^1$</td>
<td>10</td>
<td>0.0413</td>
</tr>
<tr>
<td>$A^1$</td>
<td>$B^0$</td>
<td>$C^0$</td>
<td>1</td>
<td>0.00413</td>
</tr>
<tr>
<td>$A^1$</td>
<td>$B^0$</td>
<td>$C^1$</td>
<td>10</td>
<td>0.0413</td>
</tr>
<tr>
<td>$A^1$</td>
<td>$B^1$</td>
<td>$C^0$</td>
<td>10</td>
<td>0.0413</td>
</tr>
<tr>
<td>$A^1$</td>
<td>$B^1$</td>
<td>$C^1$</td>
<td>100</td>
<td>0.413</td>
</tr>
</tbody>
</table>

Factorization in Undirected Models

- How do we define probability distributions on undirected models?
- Define a clique $C$ to be a maximal set of nodes such that each node is connected to every other node in the set.
- Define the distribution $P(S) = \frac{1}{Z} \prod_C \psi_C(S_C)$ where each $\psi_C$ is an arbitrary potential function on clique $C$, and $Z$ is a normalizer.
- Theorem: For any undirected graph $G$, any distribution which satisfies the factorization property will be globally Markov.
- **Hammersley Clifford Theorem**: For strictly positive distributions (i.e., every instantiation of variables has non-zero probability), the global Markov property is equivalent to the factorization property.
Random Field Models

- We showed above that the maximum entropy approach requires that the resulting probability distribution be in an exponential form.
- As it turns out, when we consider probability distributions on undirected graphical models, the form of the distribution is exactly in the same exponential form. For conditional models, we get

\[ P_\lambda(y|x) = \frac{1}{Z_\lambda} e^{\left( \sum_i \lambda_i f_i(x, y) \right)} \]

where \( Z_\lambda(x) = \sum_y e^{\left( \sum_i \lambda_i f_i(x, y) \right)} \)

- The features \( f_i(x, y) \) are now associated with cliques of the undirected model.

Conditional Random Field Models

- Lafferty et al [ICML ’01] define an undirected graphical model called a CRF that is similar to an HMM (if we drop the edge directions).
- Recall that in an HMM, the joint distribution \( P(x, y) \) was specified as

\[ P(q, y) = P(q_0) \prod_{t=0}^{T-1} P(q_{t+1}|q_t) \prod_{t=0}^{T} P(y_t|q_t) \]

- A CRF is a discriminant model, and so we focus mainly on modeling the conditional distribution \( P(q|y) \) instead of the joint distribution.
- One way to understand this model is that we interpret the components of the joint distributions, not as probabilities, but as arbitrary potentials.

\[ P(q|y) = \frac{P(q, y)}{P(y)} = \frac{1}{Z} \prod_{t=0}^{T-1} \phi(q_t, q_{t+1}) \prod_{t=0}^{T} \phi(q_t, y_t) \]

where the normalizer makes this a valid distribution.
Generative vs. Discriminative

- Generative approaches (e.g., HMMs) are based on modeling $P(x)$ and $P(x|y)$. Discriminative approaches directly model $P(y|x)$ and ignore $P(x)$.
- Generative models lead to directed graphical models, and discriminative models lead to undirected graphical models.
- Likelihood for generative models factors nicely for completely observed graphical models. For discriminative models, likelihood is non-factorable due to normalizing constant ($Z$).
- Generative models make it possible to handle missing data through EM; Discriminative approaches cannot deal with missing data easily.
- Although likelihood term for completely observed discriminative models is non-factorable, and requires numerical methods, it is unimodal and concave with a global optimum.
- Generative models (HMMs) have been hugely successful in speech and many other domains. Discriminative models, like CRFs, offer significant advantages in domains like text, where multiple overlapping feature sets are useful (e.g., “is-capitalized”, “is-noun” etc.). They have also been applied extensively in vision.