STAT 538

Lecture 8

Maximum Entropy Models

©Marina Meilă

mmp@stat.washington.edu

1 Entropy and KL divergence

Assume that the sample space is Ω, a (typically large) finite space. For any distribution $p : \Omega \rightarrow [0, 1]$ the **entropy** is defined as

$$H(p) \;=\; -\sum_{x\in\Omega} p(x)\log p(x) \tag{1}$$

In the above, and throughout this chapter, we adopt the convention $0 \log = 0$.

The function $H(p)$ is non-negative and concave on the space of all distributions over Ω. The minimum $H = 0$ is attained for deterministic distributions and the maximum $H_{max} = \log|\Omega|$ is attained for the uniform distribution over Ω.

The entropy measures the uncertainty in a given distribution. Closely related to the entropy is the **Kullbach-Leibler divergence** between two distributions:

$$D(p||q) \;=\; \sum_{x\in\Omega} p(x)\log\frac{p(x)}{q(x)} \tag{2}$$

The KL-divergence is asymmetric in p, q. It is non-negative, convex in (p, q) and attains the minimum $D(p||q) = 0$ iff $p \equiv q$ (proofs and details below).

The entropy and KL divergence can also be defined for continuous sample spaces, by replacing the summation with an integral. The only property that changes is that the entropy of a continuous distribution is not always non-negative. All the other properties mentioned here remain the same.

Properties of H and D

1. $H \geq 0$ (obvious)

2. H is a concave function of p (we know $x \ln x$ is convex)

3. $H(p) = 0$ iff p deterministic

4. $H(p) = \max$ for $p = u$=the uniform distribution; in this case $H(u) = \ln|\Omega|$
   Proof Let $\Omega = \{0, 1, \ldots m\}$ w.l.o.g. Then, H is a function of the m variables $p_{1:m}$, with $\sum_{x=0}^m p_x = 1$ and $p_0 = 1 - \sum_{x=1}^m p_x$.

$$H(p) \;=\; -\sum_{x=1}^m p_x \ln p_x - (1 - \sum_{x=1}^m p_x) \ln(1 - \sum_{x=1}^m p_x) \tag{3}$$

$$\frac{\partial H}{\partial p_x} \;=\; -\ln p_x - 1 + \ln(1 - \sum_{x=1}^m p_x) + 1 \;=\; 0 \tag{4}$$

$$p_x \;=\; (1 - \sum_{x'=1}^m p_{x'}) \;=\; p_0 \tag{5}$$

   In other words, all p_x must be equal.

5. $D(p||q) \geq 0$
   (D is a Bregman divergence, it corresponds to the convex function $x \ln x$; see Lecture 6)

6. $D(p||q)$ convex in (p, q)
   Proof We use the *perspective* of the convex function $-\ln x$. We have (BV 3.2.6) that $t(-\ln \frac{x}{t})$ is also convex jointly in (x, t), and we set $t = p_i, x = q_i$. It follows that $-p_i \ln \frac{q_i}{p_i} = p_i \ln \frac{p_i}{q_i}$ is convex, and therefore D is convex.

7. We define the **conditional entropy** of two random variables X, Y with joint distribution p_{XY} by

$$H(X|Y) \;=\; \sum_y p_Y(y) \sum_x H(p_{X|Y=y}) \tag{6}$$

   Thus, the conditional entropy is the average of the entropies $H(p_{X|Y=y})$. We have that

$$H(X|Y) \;=\; \sum_x \sum_y p_y \underbrace{\left(-p_{x|y} \ln p_{x|y}\right)}_{\text{concave}} \tag{7}$$

$$\leq \sum_x \left(-\sum_y p_y p_{x|y}\right)\left(\ln \sum_y p_y p_{x|y}\right) \tag{8}$$

$$= \sum_x -p_x \ln p_x \;=\; H(p_x) \;\equiv\; H(X) \tag{9}$$

In other words, $\boxed{H(X|Y) \leq H(X)}$, or conditioning (observing) another variable Y decreases the uncertainty in X. The amount of the decrease is called the **mutual information** between X and Y.

$$I(X||Y) \;=\; H(X) - H(X|Y) \;=\; H(Y) - H(Y|X) \tag{10}$$

(proving this last equality is left as an exercise)

8. The mutual information is non-negative, and is 0 iff the two variables are independent, as a consequence of the following identity

$$\boxed{I(X||Y) \;=\; D(p_{XY}||p_X p_Y)} \tag{11}$$

(proof by direct calculation).

9. For two variables X, Y with joint distribution $P_{XY} = [p_{xy}]_{x,y}$, the **joint entropy** is defined as

$$H(X,Y) \;\equiv\; H(P_{XY}) \;=\; -\sum_{xy} p_{xy} \ln p_{xy} \tag{12}$$

It is easy to verify that

$$H(X,Y) \;=\; H(X) + H(Y|X) \;\leq\; H(X) + H(Y). \tag{13}$$

The above inequality is satisfied with equality only if $X \perp Y$, i.e. only if X and Y are independent.

10. Aggregation decreases the entropy. Let the random variable Y be a deterministic function of X. When X is discrete, that means that several values x may be mapped to the same value y. Intuitively, the values x are *aggregated* into sets labeled by y. Hence, $p_y = \sum_{x \to y} p_x$.

$$H(X) \;=\; \sum_x p_x(-\ln p_x) \;=\; \sum_y \sum_{x \to y} p_x(-\ln p_x) \tag{14}$$

$$= \sum_y p_y \left[\sum_{x \to y} \frac{p_x}{p_y} \ln \frac{1}{p_x}\right] \tag{15}$$

$$\geq \; \sum_y p_y \left[\ln \sum_{x\to y} \frac{p_x}{p_y}\frac{1}{p_x}\right] \quad (\text{because } -\ln z \text{ is convex}) \tag{16}$$

$$= \; \sum_y p_y \left[\ln \frac{n_y}{p_y}\right] \quad \text{where } n_y = \#\{x \to y\} \tag{17}$$

$$= \; H(Y) + \sum_y p_y \ln n_y \;\geq\; H(Y) \tag{18}$$

11. Mixing increases entropy. For the mixture distribution $tp + (1-p)q$, by the concavity of the entropy, it follows that

$$H(tp + (1-t)q) \geq tH(p) + (1-t)H(q) \tag{19}$$

2 The Maximum Entropy Principle

Assume that we have a set of N observations $\mathcal{D} = \{x^{(1)}, \ldots x^{(N)}\} \subseteq \Omega$ from an unknown distribution p. The observation define the **empirical distribution** $\tilde{p}$

$$\tilde{p}\;(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x^{(j)}}(x) \tag{20}$$

where

$$\delta_{x^{(j)}}(x) \;=\; \begin{cases} 1 & x = x^{(j)} \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

From now on we assume that the data is represented by the empirical distribution $\tilde{p}$.

We also have a set of **features** $f_i(x)$ of the data; they are functions $f_i : \Omega \to (-\infty, \infty)$, $i = 1, \ldots K$.

The **Maximum Entropy Principle** states that the "best" model of the data is the distribution q^* representing the solution to the following problem

$$\max_q H(q) \quad \text{s.t. } E_q[f_i] \;=\; E_{\tilde{p}}\;[f_i] \text{ for all } i = 1, \ldots K. \tag{22}$$

This optimization problem has a concanve objective to be maximized (which is equivalent to minimizing the convex objective $-H(q)$), linear constraints and an overall convex space for the "variable" q. Thus, it is a convex optimization problem and we know that if a solution exists, then it is unique.

Let us now apply the usual convex optimization machinery to find a general form for the solution. The Lagrangean is

$$L(q, \lambda) \;=\; -H(q) - \sum_i \lambda_i (E_q[f_i] - E_{\tilde{p}}\,[f_i]) - \lambda_0 (\sum_{x \in \Omega} q - 1) \tag{23}$$

In the above, λ_i are the Lagrange multipliers associated to the each of the constraints $i = 1, \ldots K$, and λ_0 represents the normalization constraint on q. Note that if Ω is finite, q is a vector indexed by x. We take the partial derivative of L w.r.t each vector element $q(x)$. For this, a very useful identity will be the following:

$$\frac{d}{dy} y \log y \;=\; \log y + 1 \tag{24}$$

Therefore,

$$\frac{\partial L}{\partial q(x)} \;=\; \log q(x) + 1 - \sum_i \lambda_i f_i(x) - \lambda_0 \tag{25}$$

By equating the above with 0 we obtain

$$\begin{aligned} \log q(x) \;&=\; \sum_i \lambda_i f_i(x) + \lambda_0 - 1 & (26)\\ \text{or} & & (27)\\ q(x) \;&\propto\; e^{\sum_i \lambda_i f_i(x)} & (28)\\ \text{or} & & (29)\\ q \;&=\; \frac{1}{Z_\lambda} e^{\lambda^T f} & (30) \end{aligned}$$

The **normalization constant** Z_λ, also known as the **partition function** is defined as

$$Z_\lambda \;=\; \sum_{x \in \Omega} e^{\sum_i \lambda_i f_i(x)} \tag{31}$$

This is the general form of the solution of the Maximum Entropy problem. Note that although the formulation was non-parametric, i.e we were optimizing over all possible distributions over Ω, the resulting solution depends on K parameters, one for each constraint.

2.1 Examples

1. Without any constraints, the solution reduces to $q^* \propto 1$, i.e. the uniform distribution.

2. Let $x = (x_1, \dots x_d)$ a d-dimensional vector of binary variables $x_j \in \{0, 1\}$. Let $f_j = x_j$ for $j = 1, \dots d$. Then,

$$q(x) \propto e^{\sum_j \lambda_j x_j} = \prod_j (e^{\lambda_j})^{x_j} = \prod_j \theta_j^{x_j} \tag{32}$$

   In other words, if the features depend only on one variable, in the maximum entropy distribution the variables are independent. In particular, in our case we recover the (multivariate) binomial distribution.

3. Let us solve the above case again, explicitly, for $x = (y, z) \in \{0, 1\}^2$. Define q by its four parameters $q_{11}, q_{10}, q_{01}, q_{00}$. We have three constraints $q_{11} + q_{10} = \bar{y}$, $q_{11} + q_{01} = \bar{z}$, $q_{11} + q_{10} + q_{01} + q_{00} = 1$, where $\bar{y}, \bar{z}$ are respectively the sample means of y, z. Hence, the solution q depends on one free parameter only; let that be q_{11}. We can write q in the following way:

$$\begin{bmatrix} q_{11} & \bar{y} - q_{11} \\ \bar{z} - q_{11} & 1 - \bar{y} - \bar{z} + q_{11} \end{bmatrix}$$

   and the objective is

$$\begin{aligned} H(q_{11}) &= -q_{11} \log q_{11} - (\bar{y} - q_{11}) \log(\bar{y} - q_{11}) - (\bar{z} - q_{11}) \log(\bar{z} - q_{11}) \\ &\quad -(1 - \bar{y} - \bar{z} + q_{11}) \log(1 - \bar{y} - \bar{z} + q_{11}) \end{aligned} \tag{33}$$

   Taking the derivative we get

$$\frac{d}{dq_{11}} H = \log \frac{(\bar{y} - q_{11})(\bar{z} - q_{11})}{q_{11}(1 - \bar{y} - \bar{z} + q_{11})} = 0 \tag{34}$$

   The above is solved for $q_{11} = \bar{y}\bar{z}$ which amounts to independence between y and z.

4. Assume the same 2-dimensional binary variable sample space, but now with the feature $f(y, z) = 1$ iff $y = z$. Let the expectation of f from

the data be $\bar{f} = \frac{N_{y=z}}{N}$. Then

$$\begin{aligned}
L(q, \lambda, \lambda_0) &= \sum_{y,z} q_{yz} \log q_{yz} - \lambda(q_{00} + q_{11} - \bar{f}) - \lambda_0(\sum_{yz} q_{yz} - 1) && (35) \\
\frac{\partial L}{\partial q_{00}} &= \log q_{00} + 1 - \lambda - \lambda_0 \; = \; 0 && (36) \\
\frac{\partial L}{\partial q_{11}} &= \log q_{11} + 1 - \lambda - \lambda_0 \; = \; 0 && (37)
\end{aligned}$$

From the last two equations, we find that $q_{11} = q_{00}$ and therefore $= \bar{f}/2$. Thus, the solution q is

$$\begin{bmatrix} \frac{\bar{f}}{2} & \frac{1\bar{f}}{2} \\ \frac{1-\bar{f}}{2} & \frac{\bar{f}}{2} \end{bmatrix}$$

Again, the solution makes q as uniform as possible under the given constraints. Note that because the feature depends on both y, z the two variables are dependent in the resulting distribution.

5. **Exercise** Take the same sample space as above, with the unique feature $f(y, z) = 1$ iff $y = z = 1$. What is the corresponding MaxEnt distribution?

6. Assume now that $\Omega = (-\infty, \infty)$ and thus we have a continuous Maximum Entropy problem. Let the features be $f_1(x) = x,\; f_2(x) = x^2$. Then the maximum entropy distribution is

$$q(x) \;\propto\; e^{\lambda_1 x + \lambda_2 x^2} \tag{38}$$

which represents a Gaussian that fits the first two moments of the observed data.

7. **Exercise** Markov random fields and decomposable graphical models (aka junction trees) over discrete domains are maximum entropy distributions. Can you identify the features that they are matching?

3 Minimum Relative Entropy

The Minimum Relative Entropy (MRE) problem generalizes the Maximum Entropy (ME) problem to the case when we also have a **prior distribution**

q_0. We want to find the model that matches the **sufficient statistics** $E_{\tilde{p}}\ [f_i]$ of the data and is close to the prior.

$$\min_q D(q||q_0) \quad \text{s.t. } E_q[f_i] \ = \ E_{\tilde{p}}\ [f_i] \ \text{ for } i = 1, \ldots K \tag{39}$$

The Lagrangean of this problem is

$$L(q, \lambda) \ = \ \sum_{x\in\Omega} q(x) \log \frac{q(x)}{q_0(x)} - \sum_i \lambda_i(\sum_{x\in\Omega} f_i(x)q(x) - E_{\tilde{p}}\ [f_i]) + \lambda_0(\sum_{x\in\Omega} q(x) - 1) \tag{40}$$

And the solution is

$$\begin{aligned} q^*(x) &\propto q_0(x)e^{\sum_i \lambda_i f_i(x)} && (41)\\ \text{or} & \\ q^* &= \frac{1}{Z_\lambda} q_0 e^{\lambda^T f} && (42)\\ \text{with} & \\ Z_\lambda &= \sum_{x\in\Omega} q_0(x) e^{\sum_i \lambda_i f_i(x)} && (43) \end{aligned}$$

Note that the ME problem is a special case of the MRE problem with $q_0 \propto 1$ the uniform distribution. Indeed, it is easy to see that

$$D(q||\text{uniform}) \ = \ \sum_{\mathrm{x}\in\Omega} \mathrm{q} \log \mathrm{q} - \sum_{\mathrm{x}\in\Omega} \mathrm{q} \log \frac{1}{|\Omega|} \ = \ -\mathrm{H}(\mathrm{q}) - \log|\Omega| \tag{44}$$

3.1 The relationship with exponential family models

From (42) we have that

$$\log q(x) \ = \ \lambda^T f(x) - \log Z_\lambda + \log q_0(x) \tag{45}$$

A family of probability distributions that can be put in this form is called and **exponential family** model. Exponential family models comprise (multivariate) normal distributions, Markov random fields (with positive distributions), binomial and multinomial models, etc. They have many convenient properties, some of which are evident from the definition above. For example, exponential family models are essentially the only parametric models that have a finite number of sufficient statistics[1]; they have **conjugate priors**; from the differential geometry p.o.v, exponential families repreent **flat manifolds**, i.e affine function spaces spanned by the vectors f_i.

We also know that Maximum Likelihood (ML) estimation for exponential models consists in fitting the sufficient statistics of the data. Therefore, finding the optimal parameters λ for MRE/ME models can be seen as a ML estmation. The next section discusses this in detail.

4 The ME-ML duality

The ME/MRE problem is a convex optimization problem. Here we examine its dual and show that it represents a Maximum Likelihood problem.

The **dual objective** of problem (39) is given by $L(\lambda) = \sup_q h(q, \lambda)$. We can obtain L explicitly by substituting the solution (42) into h.

$$\begin{aligned} L(\lambda) &= h(q^* \ , \lambda) && (46) \\ &= \sum_{x\in\Omega} q^* \ \log \frac{q_0 e^{\sum_i \lambda_i f_i}}{Z_\lambda q_0} - \sum_i \lambda_i (\sum_{x\in\Omega} f_i q^* \ - E_{\tilde{p}} \ [f_i]) && (47) \\ &= \sum_{x\in\Omega} q^* \ \sum_i \lambda_i f_i - \sum_{x\in\Omega} q^* \ \log Z_\lambda - \sum_i \lambda_i \sum_{x\in\Omega} f_i q^* \ + \sum_i \lambda_i E_{\tilde{p}} \ [f_i] && (48) \\ &= \sum_i \lambda_i E_{\tilde{p}} \ [f_i]) - \log Z_\lambda && (49) \end{aligned}$$

Hence, the dual of the MRE problem is

$$\max_\lambda \lambda^T E_{\tilde{p}} \ [f]) - \log Z_\lambda \qquad (50)$$

Since all the constraints in the primal problem were equality constraints, the dual has no constraints on λ.

We now show that the dual problem represent maximixing a likelihood. The log-likelihood of the parameter set λ given a dataset $\mathcal{D}$ is

$$l(\lambda) \ = \ \sum_{x^i \in \mathcal{D}} \log q(x^i) \qquad (51)$$

[1] Distributions that are piecewise uniform may also have finite sufficient statistics. In their case, the sufficient statistics are intervals in which the data lie.

$$\begin{aligned} &= N \sum_{x\in\Omega} \tilde{p}\ (x) \log q^*(x) && (52)\\ &= N \sum_{x\in\Omega} \tilde{p}\ [\log q_0 + \sum_i \lambda_i f_i - \log Z_\lambda] && (53)\\ &= N(\sum_i \lambda_i E_{\tilde{p}}\ [f_i] - \log Z_\lambda) + \text{constant} && (54)\\ &= NL(\lambda) + \text{constant} && (55) \end{aligned}$$

Therefore, solving the MRE problem is equivalent to maximizing the likelihood of the exponential model given by (42).

It is also easy to show that maximizing the likelihood for any model family $\{q\}$ is equivalent to miniminzing the KL divergence from $\tilde{p}$ to $\{q\}$.

$$\begin{aligned} D(\tilde{p}\ \|q) &= \sum_{x\in\Omega} \tilde{p}\ \log \frac{\tilde{p}}{q} && (56)\\ &= \underbrace{-\sum_{x\in\Omega} \tilde{p}\ \log q}_{-l(q)/N} - H(\tilde{p}\ ) && (57) \end{aligned}$$

We can summarize the previous sections in the following way: for a fixed set of features f and a fixed data set we define

$$\begin{aligned} \mathcal{Q} &= \{Q \mid q \propto q_0 e^{\lambda^T f}\} && (58)\\ \mathcal{P} &= \{p \mid E_p[f] = E_{\tilde{p}}\ [f]\} && (59) \end{aligned}$$

i.e the family of exponential models with sufficient statistics f and the family of distributions that fit the sufficient statistics of the data. Let $\overline{\mathcal{Q}}$ be closure of $\mathcal{Q}$ under the Euclidean norm. Then (see Della Pietra & al) the MRE distribution q^* is the unique distribution in $\mathcal{P} \cap \overline{\mathcal{Q}}$ and is also the unique distribution satisfying

$$q^* = \text{argmin}_{\mathcal{P}} D(q\|q_0) = \text{argmin}_{\overline{\mathcal{Q}}} D(\tilde{p}\ \|q) \qquad (60)$$

The first optimization in the above represents the original MRE problem, the second is the dual ML problem. Minimizing a KL divergence to a set can be thought of as “projecting” a distribution on the respective set. Note that in our case the two projections differ since they are reversed forms of

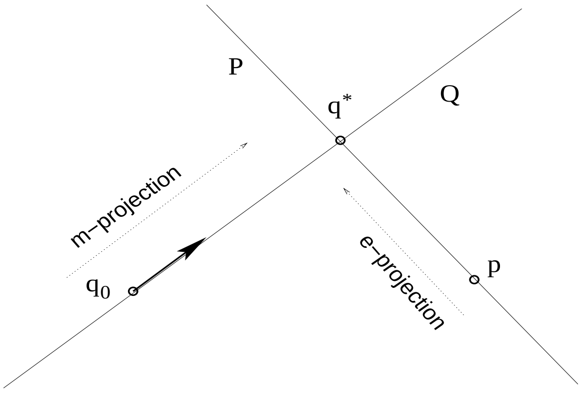

Figure 1: The duality between MRE and ML: the m-projection corresponds to ML estimation, the e-projection corresponds to MRE estimation.

the KL divergence. Thus, one obtains the same solution either by the **e-projection** of the prior q_0 on the data manifold $\mathcal{P}$ or by the **m-projection** of the empirical distribution $\tilde{p}$ on the model manifold $\overline{\mathcal{Q}}$.

The similarity to the EM algorithm is not incidental - see for example Neal and Hinton "A new view of the EM algorithm" for a view of the Expectation Maximization algorithm that emphasizes the alternating minimization of KL divergences.

A **Pytagorean theorem** is yet another equivalent way of characterizing q^* . For any $q \in \mathcal{Q}$ and any $p \in \mathcal{P}$,

$$D(p||q) \;=\; D(p||q^*\,) + D(q^*\,||q) \tag{61}$$

Proof Let $p_1, p_2 \in \mathcal{P},\; q_1, q_2 \in \mathcal{Q},\; q_2 = q_1 e^{\lambda^T f}$. Then

$$\begin{aligned} D(p_1||q_1) - D(p_1||q_2) - D(p_2||q_1) + D(p_2||q_2) \;&=\; \sum_{x\in\Omega} p_1 \log\frac{q_2}{q_1} + \sum_{x\in\Omega} p_2 \log\frac{q_2}{q_1} && (62)\\ &=\; \sum_{x\in\Omega}(p_1 - p_2)\lambda^T f && (63)\\ &=\; \lambda^T(E_{p_1}[f] - E_{p_2}[f]) && (64)\end{aligned}$$

If now $p_1 = q_1 = q^*$, $p_2 = p,\; q_2 = q$ we obtain

$$0 - D(q^*\,||q) - D(p||q^*\,) + D(p||q) \;=\; (\lambda - \lambda^*)^T(\underbrace{E_{q^*}\,[f]}_{E_{\tilde{p}}\;[f]} - \underbrace{E_p[f]}_{E_{\tilde{p}}\;[f]}) \;=\; 0 \tag{65}$$

5 Estimating the parameters

For some ME models, there is a direct relationship between the parameters and the sufficient statistics; For e.g normal distributions, multinomial distributions, decomposable models the parameter fitting is done analytically with a simple formula. For other exponential models, we have to resort to iterative methods.

Gradient ascent. One method to find the parameters of the MRE distribution is to find the unique maximum of $h(\lambda)$ using equation **??**. Here we have the problem of computing the normalization constant Z_λ which involves a summation over the whole sample space. Sometimes Z_λ and its partial derivatitives can be computed analytically; we shall see some examples in section **??**. In the other cases, one has to approximate these derivatives. The most direct method to do it is to recall that $\frac{\partial Z_\lambda}{\partial \lambda_i} = E_q[f_i]$ and apply a Markov-chain Monte Carlo method to compute the expectations.

Both **Gibbs and Metropolis sampling** from a ME distribution are straightforward since

$$\frac{q(x)}{q(x')} = \frac{q_0(x)}{q_0(x')} e^{\lambda^T (f(x) - f(x'))} \tag{66}$$

For q_0 uniform, binary features, and x, x' differring in only one feature f_j, the above ratio becomes equal to e^{λ_j}. Note also that one needs only one sample from q to estimate all the derivatives.

Improved Iterative Scaling is a method introduced by Della Pietra and al. which is generally faster than the gradient method.

IMPROVED ITERATIVE SCALING (IIS) ALGORITHM

Given $\tilde{p}$, q_0 and the features $f = (f_1, \ldots f_K)$. Assume q_0 is absolutely continuous w.r.t $\tilde{p}$ (i.e the prior gives non-zero probability to the data). Denote

$$f_\#(x) = \sum_i f_i(x) \tag{67}$$

1. Initialize $q \leftarrow q_0, \ \lambda \leftarrow 0$

2. For each i, let γ_i be the unique solution of

$$E_q[f_i e^{\gamma_i f_\#}] \;=\; E_{\tilde{p}}\;[f_i] \tag{68}$$

3. Set $q \leftarrow q e^{\sum_i \gamma_i f_i}$ or, equivalently, set $\lambda \leftarrow \lambda + \gamma$

4. If q has not converged yet, go to step 2

The γ estimation in step 2 of the algorithm can be done by bisection noting that the function is monotonic in γ.

6 Learning the features

Della Pietra et al. present a method for learning features from data. The method is a greedy algorithm, that consists of two steps: first, from a set of **candidate features** C one feature is selected to be added to the model; second, the optimal parameters λ of the augmented model are fit to the data.

Field Induction Algorithm

Input: $\tilde{p}$, q_0, a rule for constructing candidate features or a set of possible features

1. Set $q \leftarrow q_0$

2. **Feature selection**

   (a) Construct the current set of candidate features C

   (b) For $g \in C$ compute the gain $G(g) = \max_\alpha D(\tilde{p}\,||q) - D(\tilde{p}\,||qe^{\alpha g}/Z_\alpha)$ and $\hat{\alpha}$ that maximizes the gain

   (c) Add $g = \text{argmax}_C G(g)$ to the current set of features

3. **Parameter fitting** Refit the model parameters by IIS, starting from the previous parameter set λ and $\hat{\alpha}$.

4. Return to step 2

The algorithm is guaranteed to improve the likelihood of the data at each iteration, and to give the best model for the given features, but it incorporates no model selection method. So the stopping criterion and model validation are left to the standard methods.

Below we discuss the feature selection in more detail.

6.1 Greedy feature selection

In the following we assume that all features take values in $\{0,1\}$. Assume that q is the current ME distribution and $g \in C$ is a new feature. Let $Z = \sum_{x\in\Omega} qe^{\alpha g}$, $\tilde{q} = qe^{\alpha g}/Z$ and $G(\alpha, g) = D(\tilde{p}\ ||q) - D(\tilde{p}\ ||\tilde{q})$ the gain of adding g with parameter α to the model. Then, it is easy to show the following three equalities:

$$\begin{aligned} G(\alpha, g) &= \alpha E_{\tilde{p}}\ [g] - \log E_q[e^{\alpha g}] && (69) \\ \frac{\partial}{\partial\alpha} G(\alpha, g) &= E_{\tilde{p}}\ [g] - E_{\tilde{q}}[g] && (70) \\ \frac{\partial^2}{\partial\alpha^2} G(\alpha, g) &= -Var_{\tilde{q}}[g]\ \le\ 0 && (71) \end{aligned}$$

This shows that $G(\alpha, g)$ has a unique maximum w.r.t α.

Lemma We now show that the maximum is at

$$\hat{\alpha} \;=\; \log \frac{\tilde{p}\ (g=1)(1-q(g=1))}{(1-\tilde{p}\ (g=1))q(g=1)} \qquad (72)$$

We have that

$$\tilde{q} \;=\; \begin{cases} \frac{qe^{\alpha}}{Z} & \text{if } g=1 \\ \frac{q}{Z} & \text{if } g=0 \end{cases} \qquad (73)$$

Therefore, $E_{\tilde{q}}[g] = e^{\alpha}q(g=1)/Z$ and $1 - E_{\tilde{q}}[g] = q(g=0)/Z = (1-q(g=1))/Z$. Remembering from (70) that $E_{\tilde{q}}[g] = E_{\tilde{p}}\ [g]$ at the optimal α, we obtain

$$\begin{aligned} E_{\tilde{p}}\ [g] &= e^{\alpha}q(g=1)/Z && (74) \\ 1 - E_{\tilde{p}}\ [g] &= (1-q(g=1))/Z && (75) \end{aligned}$$

By taking the ratio of the two, we obtain 72.

Lemma Denote by $\underline{p}, \underline{q}$ the probabilities $\tilde{p}\ (g = 1),\ q(g = 1)$ respectively. Then, the gain $G(g)$ is equal to $D(B_{\underline{p}}||B_{\underline{q}})$ where $B_{\underline{p}}, B_{\underline{q}}$ are respectively the Bernoulli distributions with $\underline{p}, \underline{q}$ as probabilities of success.

Proof

$$\begin{aligned} D(B_{\underline{p}}||B_{\underline{q}}) &= \underline{p}\log\frac{\underline{p}}{\underline{q}} + (1-\underline{p})\log\frac{1-\underline{p}}{1-\underline{q}} && (76)\\ &= \underline{p}\log\frac{\underline{p}(1-\underline{q})}{\underline{q}(1-\underline{p})} - \log\frac{1-\underline{q}}{1-\underline{p}} && (77)\end{aligned}$$

By 69,

$$\begin{aligned} G(\hat{\alpha}, g) &= \underline{p}\log\frac{\underline{p}(1-\underline{q})}{\underline{q}(1-\underline{p})} - \log E_q[e^{\alpha g}] && (78)\\ &= \underline{p}\log\frac{\underline{p}(1-\underline{q})}{\underline{q}(1-\underline{p})} - \log[e^{\alpha}q(g=1) + 1.q(g=0)] && (79)\\ &= \underline{p}\log\frac{\underline{p}(1-\underline{q})}{\underline{q}(1-\underline{p})} - \log\left[\frac{\underline{p}(1-\underline{q})}{\underline{q}(1-\underline{p})}\underline{q} + 1 - \underline{q}\right] && (80)\\ &= \underline{p}\log\frac{\underline{p}(1-\underline{q})}{\underline{q}(1-\underline{p})} - \log\frac{1-\underline{q}}{1-\underline{p}} && (81)\end{aligned}$$

Thus, one adds the feature g along which the discrepancy of p and q is maximized.

7 Maximum Entropy Discrimination

7.1 The ideea

Here we apply the ME framework to classification problems. We have a dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)\}$ of N labeled examples and a family of classifiers $\mathcal{F} = \{f(.;\theta)\}$ parametrized by θ. The values of the labels are in $\{\pm 1\}$ and the label of x given by $f(.;\theta)$ is $\operatorname{sign} f(x,;\theta)$.

Standard classification chooses an $f \in \mathcal{F}$ that has low classification error and perhaps obeys other regularization conditions. In the **Maximum Entropy** approach to classification, we find a distribution q over $\mathcal{F}$, such that the expected value f^* of all classifiers $f \in \mathcal{F}$ under q classifies the training set well. Naturally, more than one such q exists and we choose the one which has maximum entropy.

$$\min_q -H(q) \quad \text{s.t. } y_i E_q[f(x_i)] \geq 1 \text{ for } i = 1, \ldots N \tag{82}$$

A direct generalization of ME classification is to introduce a prior q_0 over $\mathcal{F}$ and then to choose the q that is nearest to the prior. This is the MRE discrimination problem.

$$\min_q D(q||q_0) \quad \text{s.t. } y_i E_q[f(x_i)] \geq 1 \text{ for } i = 1, \ldots N \tag{83}$$

7.2 The MRE solution

The problems 82, 83 are convex problems with linear constraints; by applying the usual transformations we obtain the Lagrangean, the general exponential form of the solution and the dual. Note that unlike the unsupervised ME problem, the current problem has inequality constraints, so that the Lagrange multipliers will have to be ≥ 0. Another difference is that the sums are now replaced with integrals, as the function families we typically consider are continuous.

$$\begin{aligned}
h(q, \lambda) &= D(q||q_0) - \sum_i \lambda_i (y_i E_q[f(x_i)] - 1) \\
\frac{\partial h}{\partial q(f)} &= \log q(f) - \log q_0(f) - \sum_i \lambda_i y_i f(x_i) \\
q^*(f) &= \frac{1}{Z_\lambda} q_0(\theta) e^{\sum_i \lambda_i y_i f(x_i)} \\
Z_\lambda &= \int q_0(\theta) e^{\sum_i \lambda_i y_i f(x_i;\theta)} d\theta \\
L(\lambda) &= h(q^*(\lambda), \lambda) \\
&= \int \frac{1}{Z_\lambda} q_0(\theta) e^{\sum_i \lambda_i y_i f(x_i;\theta)} [\sum_i \lambda_i y_i f(x_i;\theta) - \log q_0(\theta) - \log Z_\lambda] d\theta - \sum_i \lambda_i \Big[y_i \int \frac{1}{Z_\lambda} q_0(\theta) \\
&= \sum_i \lambda_i - \log Z_\lambda
\end{aligned}$$

Hence, the solution has again an exponential form, with one factor for each training example. The dual problem is

$$\max_{\lambda} L(\lambda) \quad \text{s.t. } \lambda_i \geq 0 \ \text{ for } i = 1, \ldots N \tag{90}$$

7.3 Computing the solution

The optimal λ's can be found by gradient ascent on the dual objective L. As previously seen, these are equal to the primal constraints.

$$\begin{aligned} \frac{\partial \log Z_\lambda}{\partial \lambda_i} &= \frac{1}{Z_\lambda} \int q_0(\theta) e^{\sum_i \lambda_i y_i f(x_i;\theta)} y_i f(x_i; \theta) d\theta \;=\; y_i E_q[f(x_i; .)] && (91) \\ \frac{\partial L}{\partial \lambda_i} &= 1 - y_i E_q[f(x_i; .)] && (92) \end{aligned}$$

In practice, we start with $\lambda_i = 0$ for all i and update the Lagrange multipliers by

$$\lambda_i \;\leftarrow\; \max[\,0,\, \lambda_i + \eta \frac{\partial L}{\partial \lambda_i}\,] \tag{93}$$

where η is a step size. This iteration will converge to the unique solution of 83 if one exists. Note that during the iteration, L increases, while the primal objective increases too. This is because our method is eseentially an exterior point method: we start from the unconstrained optimum and adjust the parameters λ until all the constraints are satisfied.

As it is usual with constrained optimization, only the λ_i parameters that correspond to tight constraints are non-zero. The corresponding x_i examples represent **support vectors** for this problem. Thus the solution is often sparse.

The ME classifier is given by

$$f^*(x) \;=\; \int q^*(\theta) f(x; \theta) d\theta \tag{94}$$

Note that f^* may not belong to the original family $\mathcal{F}$.

8 ME discrimination extensions

8.1 Using generative models

One important use of the ME classification framework is to combine **discriminative** classification, i.e classification via optimization of the decision boundaries, with **generative models** i.e probabilistic models of how the data were generated. The former have the advantage that they optimize the "right" criterion, the latter are much better at incorporating domain features.

Assume that we have a family of probabilisitic models describing the data in each class; let them be $P(x|\theta_+)$, $P(x|\theta_-)$ respectively. The models may belong to different families (i.e gaussian for the "+" class and uniform on a rectangle for the "-" class), and the parameters $\theta_\pm$ may belong to different spaces. A Likelihood ratio classifier in with these model families is

$$f(x;\theta_+,\theta_-,b) \;=\; \log\frac{P(x|\theta_+)}{P(x|\theta_-)}+b \tag{95}$$

We construct a MRE distribution over this classifier family that has the form

$$q(\theta_+,\theta_,b) \;\propto\; q_0(\theta_+,\theta_,b)e^{\sum_i\lambda_i[y_i(\log\frac{P(x_i|\theta_+)}{P(x_i|\theta_-)}+b)-1]} \tag{96}$$

If the prior q_0 factors into $q_0(\theta_+)q_0(\theta_-)q_0(b)$ then the MRE distribution also factors into independent distributions for the 3 parameters.

$$q(\theta_+,\theta_,b) \;\propto\; q_0(\theta+)e^{\sum_i\lambda_i y_i\log P(x_i|\theta_+)}q_0(\theta_-)e^{\sum_i\lambda_i y_i\log P(x_i|\theta_-)}q_0(b)e^{b\sum_i\lambda_i y_i}e^{-\sum_i\lambda_i} \tag{97}$$

There are several model classes for which computing the partition function and the required expectations can be done in closed form. For example, if $P(.|\theta_\pm)$ is an exponential family model and q_0 is the corresponding conjugate prior. In particular, graphical models with fixed structure and exponential family distributions are also tractable; in the special case of tree graphical models, one can construct more interesting graphical models that can be integrated over structures and parameters (see Jaakkola et al.).

8.2 Inseparable data (soft margins)

If the data are not separable by any model in $\mathcal{F}$, then the constraints of **??** cannot be satisfied for any q and the parameters λ tend to infinity. Now we extend the MRE framework to deal with this case. We will fix a variable margin γ_i for each example (x_i, y_i) and we will estimate a joint MRE distribution over the classfier parameters θ and the margins γ. Note once again that, if the prior q_0 is factored w.r.t θ, γ then the θ and γ will be independent under the final solution.

The **soft margin MRE problem** is

$$\min_q D(q||q_0) \quad \text{s.t. } E_q[y_i f(x_i) - \gamma_i] \ \geq \ 0 \quad \text{for } i = 1, \dots N \tag{98}$$

where q, q_0 are now distributions over $\theta, \gamma_1, \dots \gamma_N$. If $q_0(\theta, \gamma) = q_0(\theta) \prod_i q_0(\gamma_i)$ then the MRE solution is

$$q \ \propto \ \left(q_0(\theta) e^{\sum_i \lambda_i y_i f(x_i;\theta)} \right) \prod_{i=1}^{N} \left(q_0(\gamma_i) e^{-\lambda_i \gamma_i} \right) \tag{99}$$

If we denote the corresponding normalization constants by Z_θ, Z_{γ_i} (keeping in mind that they are functions of λ), then the dual objective can be written as

$$L(\lambda) \ = \ -\log Z_\theta - \sum_i \log Z_{\gamma_i} \tag{100}$$

The form of the MRE distribution for γ suggests an exponential prior

$$q_0(\gamma) \ = \ c e^{c(\gamma - 1)} \tag{101}$$

For this q_0 the MRE distribution is

$$q(\gamma) \ = \ (c - \lambda) e^{(c-\lambda)(\gamma - 1)} \tag{102}$$

In the above λ needs to be $\leq c$ for a proper q to exist. Hence, introducing a soft margin in this way is equivalent to puttig and upper bound on λ, very much like in the case of SVM's.

Note also that $E_{q_0}[\gamma_i] = 1 - 1/c$. Therefore, if $E_q[f(x_i)] \geq 1 - 1/c$ then $\lambda_i = 0$. Since $Z_{\gamma_i} = 1/(c - \lambda_i)$ each term $-\log Z_{\gamma_i}$ in the dual objective introduces a penalty $\log(c - \lambda_i)$. If λ_i approaches c from below, this penalty tends to $-\infty$, effectively stopping λ_i from growing too much.

8.3 Relationship to SVM's

The following theorem, from Jaakkola et al, shows that there is a strong relationship between SVM's and MRE.

Theorem Assume $f(x;\theta,b) = \theta^T x - b$ and $q_0(\theta,b,\gamma) = q_0(\theta)q_0(b)\prod_i q_0(\gamma_i)$ where $q_0(\theta) = N(0,I)$, $q_0(b)$ approaches a non-informative prior, and $q_0(\gamma_i)$ is given by equation 101. Then, the Lagrange multipliers are obtained by maximizing $L(\lambda)$ subject to $0 \leq \lambda \leq c$ and $\sum_{i=1}^N \lambda_i y_i = 0$ where

$$L(\lambda) = \sum_i (\lambda_i + \log(c - \lambda_i)] - \frac{1}{2}\sum_{i,j=1}^N \lambda_i \lambda_j y_i y_j x_i^T x_j \qquad (103)$$

Note the similarity between this objective and the SVM with slack variables. The only difference is the addional term $\log(c - \lambda_i)$. Moreover, for the case of separable data and $c \to \infty$ the same classifiers are obtained.