67577 – Intro. to Machine Learning **Fall semester, 2008/9**

Lecture 3: Maximum Likelihood/ Maximum Entropy Duality

Lecturer: Amnon Shashua *Scribe: Amnon Shashua* [1]

In the previous lecture we defined the principle of Maximum Likelihood (ML): suppose we have random variables $X_1, ..., X_n$ form a random sample from a discrete distribution whose joint probability distribution is $P(\mathbf{x} \mid \phi)$ where $\mathbf{x} = (x_1, ..., x_n)$ is a vector in the sample and ϕ is a parameter from some parameter space (which could be a discrete set of values — say class membership). When $P(\mathbf{x} \mid \phi)$ is considered as a function of ϕ it is called the *likelihood function*. The ML principle is to select the value of ϕ that maximizes the likelihood function over the observations (training set) $\mathbf{x}_1, ..., \mathbf{x}_m$. If the observations are sampled i.i.d. (a common, not always valid, assumption), then the ML principle is to maximize:

$$\phi^* = \text{argmax}_\phi \prod_{i=1}^m P(\mathbf{x}_i \mid \phi) = \text{argmax} \log \prod_{i=1}^m P(\mathbf{x}_i \mid \phi) = \text{argmax} \sum_{i=1}^m \log P(\mathbf{x}_i \mid \phi)$$

which due to the product nature of the problem it becomes more convenient to maximize the log likelihood. We will take a closer look today at the ML principle by introducing a key element known as the *relative entropy* measure between distributions.

3.1 ML and Empirical Distribution

The ML principle states that the empirical distribution of an i.i.d. sequence of examples is the closest possible (in terms of relative entropy which would be defined later) to the true distribution. To make this statement clear let $\mathcal{X}$ be a set of symbols $\{a_1, ..., a_n\}$ and let $P(a \mid \theta)$ be the probability (belonging to a parametric family with parameter θ) of drawing a symbol $a \in \mathcal{X}$. Let $x_1, ..., x_m$ be a sequence of symbols drawn i.i.d. according to P. The *occurrence frequency* $f(a)$ measures the number of draws of the symbol a:

$$f(a) = |\{i \ : \ x_i = a\}|,$$

and let the *empirical distribution* be defined by

$$\hat{P}(a) = \frac{1}{\sum_{\alpha \in \mathcal{X}} f(\alpha)} f(a) = \frac{1}{\|f\|_1} f(a) = (1/m) f(a).$$

The joint probability $P(x_1, ..., x_m \mid \phi)$ is equal to the product $\prod_i P(x_i \mid \phi)$ which according to the definitions above is equal to:

$$P(x_1, ..., x_m \mid \phi) = \prod_{i=1}^m p(x_i \mid \theta) = \prod_{a \in \mathcal{X}} P(a \mid \phi)^{f(a)}.$$

[1]

The ML principle is therefore equivalent to the optimization problem:

$$\max_{P\in Q} \prod_{a\in\mathcal{X}} P(a \mid \phi)^{f(a)} \tag{3.1}$$

where $Q = \{\mathbf{q} \in R^n : \mathbf{q} \geq 0,\ \sum_i q_i = 1\}$ denote the set of n-dimensional probability vectors ("probability simplex"). Let p_i stand for $P(a_i \mid \phi)$ and f_i stand for $f(a_i)$. Since $\text{argmax}_x z(x) = \text{argmax}_x \ln z(x)$ and given that $\ln \prod_i p_i^{f_i} = \sum_i f_i \ln p_i$ the solution to this problem can be found by setting the partial derivative of the Lagrangian to zero:

$$L(\mathbf{p}, \lambda, \mu) = \sum_{i=1}^{n} f_i \ln p_i - \lambda(\sum_i p_i - 1) - \sum_i \mu_i p_i,$$

where λ is the Lagrange multiplier associated with the equality constraint $\sum_i p_i - 1 = 0$ and $\mu_i \geq 0$ are the Lagrange multipliers associated with the inequality constraints $p_i \geq 0$. We also have the complementary slackness condition that sets $\mu_i = 0$ if $p_i > 0$.

After setting the partial derivative with respect to p_i to zero we get:

$$p_i = \frac{1}{\lambda + \mu_i} f_i.$$

Assume for now that $f_i > 0$ for $i = 1, ..., n$. Then from complementary slackness we must have $\mu_i = 0$ (because $p_i > 0$). We are left therefore with the result $p_i = (1/\lambda) f_i$. Following the constraint $\sum_i p_1 = 1$ we obtain $\lambda = \sum_i f_i$. As a result we obtain: $P(a \mid \phi) = \hat{P}(a)$. In case $f_i = 0$ we could use the convention $0 \ln 0 = 0$ and from continuity arrive to $p_i = 0$.

We have arrived to the following theorem:

Theorem 1 *The empirical distribution estimate $\hat{P}$ is the unique Maximum Likelihood estimate of the probability model Q on the occurrence frequency $f()$.*

This seems like an obvious result but it actually runs deep because the result holds for a very particular (and non-intuitive at first glance) distance measure between non-negative vectors. Let $dist(\mathbf{f}, \mathbf{p})$ be some distance measure between the two vectors. The result above states that:

$$\hat{P} = \text{argmin}_{\mathbf{p}} dist(\mathbf{f}, \mathbf{p}) \quad s.t. \quad \mathbf{p} \geq 0,\ \sum_i p_i = 1, \tag{3.2}$$

for some (family?) of distance measures $dist()$. It turns out that there is only one[2] such distance measure, known as the relative-entropy, which satisfies the ML result stated above.

3.2 Relative Entropy

The relative-entropy (RE) measure $D(\mathbf{x}||\mathbf{y})$ between two non-negative vectors $\mathbf{x}, \mathbf{y} \in R^n$ is defined as:

$$D(\mathbf{x}||\mathbf{y}) = \sum_{i=1}^{n} x_i \ln \frac{x_i}{y_i} - \sum_i x_i + \sum_i y_i.$$

[2]not exactly — the picture is a bit more complex. Csiszar's 1972 measures: $dist(\mathbf{p}, \mathbf{f}) = \sum_i f_i \phi(p_i/f_i)$ will satisfy eqn. 3.2 provided that ϕ'^{-1} is an exponential. However, $dist(\mathbf{f}, \mathbf{p})$ (parameters positions are switched) will not do it, whereas the relative entropy will satisfy eqn. 3.2 regardless of the order of the parameters $\mathbf{p}, \mathbf{f}$.

In the definition we use the convention that $0 \ln \frac{0}{0} = 0$ and based on continuity that $0 \ln \frac{0}{y} = 0$ and $x \ln \frac{x}{0} = \infty$. When $\mathbf{x}, \mathbf{y}$ are also probability vectors, i.e., belong to Q, then $D(\mathbf{x}||\mathbf{y}) = \sum_i x_i \ln \frac{x_i}{y_i}$ is also known as the Kullback-Leibler divergence. The RE measure is not a distance metric as it is not symmetric, $D(\mathbf{x}||\mathbf{y}) \neq D(\mathbf{y}||\mathbf{x})$, and does not satisfy the triangle inequality. Nevertheless, it has several interesting properties which make it a fundamental measure in statistical inference.

The relative entropy is always non-negative and is zero if and only if $\mathbf{x} = \mathbf{y}$. This comes about from the log-sum inequality:

$$\sum_i x_i \ln \frac{x_i}{y_i} \geq (\sum_i x_i) \ln \frac{\sum_i x_i}{\sum_i y_i}$$

Thus,

$$D(\mathbf{x}||\mathbf{y}) \geq (\sum_i x_i) \ln \frac{\sum_i x_i}{\sum_i y_i} - \sum_i x_i + \sum_i y_i = \bar{x} \ln \frac{\bar{x}}{\bar{y}} - \bar{x} + \bar{y}$$

But $a \ln(a/b) \geq a - b$ for $a, b \geq 0$ iff $\ln(a/b) \geq 1 - (b/a)$ which follows from the inequality $\ln(x+1) > x/(x+1)$ (which holds for $x > -1$ and $x \neq 0$). We can state the following theorem:

Theorem 2 *Let $\mathbf{f} \geq 0$ be the occurrence frequency on a training sample. $\hat{P} \in Q$ is a ML estimate iff*

$$\hat{P} = \mathit{argmin}_{\mathbf{p}} D(\mathbf{f}||\mathbf{p}) \quad s.t. \quad \mathbf{p} \geq 0, \ \sum_i p_i = 1.$$

Proof:

$$D(\mathbf{f}||\mathbf{p}) = -\sum_i f_i \ln p_i + \sum_i f_i \ln f_i - \sum_i f_i + 1,$$

and

$$\text{argmin}_{\mathbf{p}} D(\mathbf{f}||\mathbf{p}) = \text{argmax}_{\mathbf{p}} \sum_i f_i \ln p_i = \text{argmax}_{\mathbf{p}} \ln \prod_i p_i^{f_i}.$$

□

There are two (related) interesting points to make here. First, from the proof of Thm. 1 we observe that the non-negativity constraint $\mathbf{p} \geq 0$ need not be enforced - as long as $\mathbf{f} \geq 0$ (which holds by definition) the closest $\mathbf{p}$ to $\mathbf{f}$ under the constraint $\sum_i p_i = 1$ *must come out non-negative.* Second, the fact that the closest point $\mathbf{p}$ to $\mathbf{f}$ comes out as a scaling of $\mathbf{f}$ (which is by definition the empirical distribution $\hat{P}$) arises because of the relative-entropy measure. For example, if we had used a least-squares distance measure $\|\mathbf{f}-\mathbf{p}\|^2$ the result would not be a scaling of $\mathbf{f}$. In other words, we are looking for a projection of the vector $\mathbf{f}$ onto the probability simplex, i.e., the intersection of the hyperplane $\mathbf{x}^\top \mathbf{1} = 1$ and the non-negative orthant $\mathbf{x} \geq 0$. Under relative-entropy the projection is simply a scaling of $\mathbf{f}$ (and this is why we do not need to enforce non-negativity). Under least-sqaures, a projection onto the hyper-plane $\mathbf{x}^\top \mathbf{1} = 1$ could take us out of the non-negative orthant (see Fig. 3.1 for illustration). So, relative-entropy is special in that regard — it not only provides the ML estimate, but also simplifies the optimization process[3] (something which would be more noticeable when we handle a latent class model next lecture).

3.3 Maximum Entropy and Duality ML/MaxEnt

The relative-entropy measure is not symmetric thus we expect different outcomes of the optimization $\min_x D(x||y)$ compared to $\min_y D(x||y)$. The latter of the two, i.e., $\min_{P \in \mathcal{Q}} D(P_0||P)$, where

[3]The fact that non-negativity "comes for free" does not apply for all class (distribution) models. This point would be refined in the next lecture.

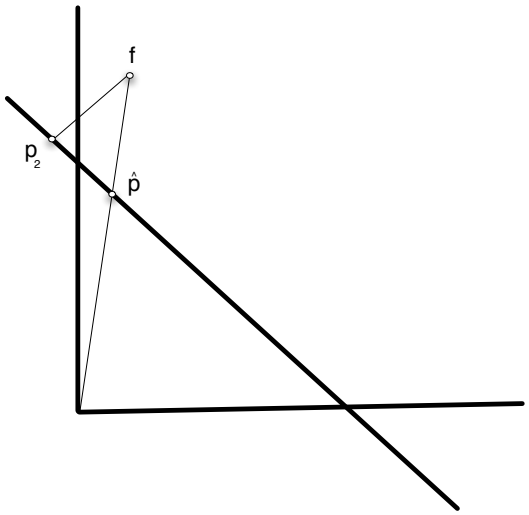

Figure 3.1: Projection of a non-neagtaive vector $\mathbf{f}$ onto the hyperplane $\sum_i x_i - 1 = 0$. Under relative-entropy the projection $\hat{P}$ is a scaling of $\mathbf{f}$ (and thus lives in the probability simplex). Under least-squares the projection p_2 lives outside of the probability simplex, i.e., could have negative coordinates.

P_0 is some empirical evidence and $\mathcal{Q}$ is some model, provides the ML estimation. For example, in the next lecture we will consider $\mathcal{Q}$ the set of low-rank joint distributions (called latent class model) and see how the ML (via relative-entropy minimization) solution can be found.

Let $H(\mathbf{p}) = -\sum_i p_i \ln p_i$ denote the *entropy* function. With regard to $\min_x D(x||y)$ we can state the following observation:

Claim 1

$$argmin_{\mathbf{p}\in\mathcal{Q}} D(\mathbf{p}||\frac{1}{n}\mathbf{1}) = argmax_{\mathbf{p}\in\mathcal{Q}} H(\mathbf{p}).$$

Proof:

$$D(\mathbf{p}||\frac{1}{n}\mathbf{1}) = \sum_i p_i \ln p_i + (\sum_i p_i)\ln(n) = \ln(n) - H(\mathbf{p}),$$

which follows from the condition $\sum_i p_i = 1$. □

In other words, the closest distribution to uniform is achieved by maximizing the entropy. To make this interesting we need to add constraints. Consider a linear constraint on $\mathbf{p}$ such as $\sum_i \alpha_i p_i = \beta$. To be concrete, consider a die with six faces thrown many times and we wish to estimate the probabilities $p_1, ..., p_6$ given only the *average* $\sum_i i p_i$. Say, the average is 3.5 which is what one would expect from an unbiased die. The *Laplace's principle of insufficient reasoning* calls for assuming uniformity unless there is additional information (a controversial assumption in some cases). In other words, if we have no information except that each $p_i \geq 0$ and that $\sum_i p_i = 1$ we should choose the uniform distribution since we have no reason to choose any other distribution. Thus, employing Laplace's principle we would say that if the average is 3.5 then the most "likely" distribution is the uniform. What if $\beta = 4.2$? This kind of problem can be stated as an optimization problem:

$$\max_{\mathbf{p}} H(\mathbf{p}) \quad s.t., \sum_i p_i = 1, \ \sum_i \alpha_i p_i = \beta,$$

where $\alpha_i = i$ and $\beta = 4.2$. We have now two constraints and with the aid of Lagrange multipliers we can arrive to the result:

$$p_i = \exp^{-(1-\lambda)} \exp^{\mu\alpha_i}.$$

Note that because of the exponential $p_i \geq 0$ and again "non-negativity comes for free"[4]. Following the constraint $\sum_i p_i = 1$ we get $\exp^{-(1-\lambda)} = 1/\sum_i \exp^{\mu\alpha_i}$ from which obtain:

$$p_i = \frac{1}{Z} \exp^{\mu\alpha_i},$$

where Z (a function of μ) is a normalization factor and μ needs to be set by using β (see later). There is nothing special about the uniform distribution, thus we could be seeking a probability vector $\mathbf{p}$ as close as possible to some prior probability $\mathbf{p}_0$ under the constraints above:

$$\min_{\mathbf{p}} D(\mathbf{p}||\mathbf{p}_0) \quad s.t., \sum_i p_i = 1, \ \sum_i \alpha_i p_i = \beta,$$

with the result:

$$p_i = \frac{1}{Z} p_{0i} \exp^{\mu\alpha_i}.$$

We could also consider adding more linear constraints on $\mathbf{p}$ of the form: $\sum_i f_{ij} p_i = b_j$, $j = 1, ..., k$. The result would be:

$$p_i = \frac{1}{Z} p_{0i} \exp^{\sum_{j=1}^{k} \mu_j f_{ij}}.$$

Probability distributions of this form are called *Gibbs Distributions*. In practical applications the linear constraints on $\mathbf{p}$ could arise from *average* information about the system such as temperature of a fluid (where p_i are the probabilities of the particles moving at various velocities), rainfall data or general environmental data (where p_i represent the probability of finding animal colonies at discrete locations in a 3D map). A constraint of the form $\sum_i f_{ij} p_i = b_j$ states that the expectation $E_p[f_j]$ should be equal to the empirical distribution $\beta = E_{\hat{P}}[f_j]$ where $\hat{P}$ is either uniform or given as input. Let

$$\mathcal{P} = \{\mathbf{p} \in R^n \ : \ \mathbf{p} \geq 0, \ \sum_i p_i = 1, \ E_p[f_j] = E_{\hat{p}}[f_j], j = 1, ..., k\},$$

and

$$\mathcal{Q} = \{\mathbf{q} \in R^n \ ; \ \mathbf{q} \text{ is a Gibbs distribution}\}$$

We could therefore consider looking for the ML solution for the parameters $\mu_1, ..., \mu_k$ of the Gibbs distribution:

$$\min_{\mathbf{q} \in \mathcal{Q}} D(\hat{\mathbf{p}}||\mathbf{q}),$$

where if $\hat{\mathbf{p}}$ is uniform then $\min D(\hat{\mathbf{p}}||\mathbf{q})$ can be replaced by $\max \sum_i \ln q_i$ (because $D((1/n)\mathbf{1}||\mathbf{x}) = -\ln(n) - \sum_i \ln x_i$).

As it turns out, the MaxEnt and ML are duals of each other and the intersection of the two sets $\mathcal{P} \cap \mathcal{Q}$ contains only a single point which *solves both problems*.

Theorem 3 *The following are equivalent:*

- *MaxEnt:* $\mathbf{q}^* = argmin_{\mathbf{p} \in \mathcal{P}} D(\mathbf{p}||\mathbf{p}_0)$

[4] Any measure of the class $dist(\mathbf{p}, \mathbf{p}_0) = \sum_i p_{0i} \phi(p_i/p_{0i})$ minimized under linear constraints will satisfy the result of $p_i \geq 0$ provided that ϕ'^{-1} is an exponential.

- *ML:* $\mathbf{q}^* = argmin_{\mathbf{q}\in\mathcal{Q}} D(\hat{\mathbf{p}}||\mathbf{q})$
- $\mathbf{q}^* \in \mathcal{P} \cap \mathcal{Q}$

In practice, the duality theorem is used to recover the parameters of the Gibbs distribution using the ML route (second line in the theorem above) — the algorithm for doing so is known as the *iterative scaling algorithm* (which we will not get into).