Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

15 Modeling as Inference

Having seen how information measures naturally arise in inference from a Bayesian perspective, we now develop the corresponding role they play in inference from a nonBayesian perspective by considering the central problem of *modeling*. In the process, our development reveals how the two perspectives are fundamentally intertwined.

As in our discussion of estimation, the nonBayesian scenario of interest is that we have some observations of a random variable y in some alphabet $\mathcal{Y}$ whose distribution lies in the class

$$\mathcal{P} = \{p_{\mathsf{y}}(\cdot; x) \in \mathcal{P}^{\mathcal{Y}} \colon x \in \mathcal{X}\}, \tag{1}$$

but is otherwise unknown. However, the problem of modeling is in general distinct from that of estimation, even when we restrict our attention to such *parametric* models. In particular, while the goal of estimation is to try to determine the unknown parameter x, the goal of modeling is to determine a good approximation $q(\cdot)$ to the true but unknown distribution for y. As in the Bayesian discussion, we restrict our attention to the case of discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$ for our initial development.

We begin with an example.

Example 1. Let $\mathsf{y}_1, \mathsf{y}_2, \dots, \mathsf{y}_{n-1}$ be i.i.d. Bernoulli random variables (over the alphabet $\mathcal{Y} = \{0,1\}$) with parameter $x = \mathbb{P}(\mathsf{y}_i = 1) \in \mathcal{X} = [0,1]$, i.e., $\mathsf{y}_i \sim \mathtt{B}(x)$, where $x \in \mathcal{X}$ is unknown. Having observed $y_1, y_2, \dots, y_{n-1}$, how should we best characterize uncertainty in y_n? In other words, what is our best estimate for the distribution of y_n?

An obvious heuristic approach to such a modeling goal would involve estimating x from the available data, then using the estimate as if it were the true parameter. In this example, we could construct the maximum likelihood estimator

$$\hat{x}_{\mathrm{ML}}^{(n-1)} = \arg\max_x p_{\mathsf{y}_1,\dots,\mathsf{y}_{n-1}}(y_1,\dots,y_{n-1}; x) = \frac{1}{n-1}\sum_{i=1}^{n-1} y_i = \frac{n_1}{n-1}, \tag{2}$$

where we recognize n_1 as the number of observations that are equal to one. We could then use

$$q(y_n) = p_{\mathsf{y}}\left(y_n; \hat{x}_{\mathrm{ML}}^{(n-1)}\right) \tag{3}$$

to approximate the distribution of y_n. As we will see later, this approach can sometimes be justified asymptotically (as $n \to \infty$). But it is not the best approach for any finite value of n, as we will make precise here.

The key challenge in modeling problems is that we need to model the distribution of y with a "good" distribution $q(\cdot)$ that does not depend on the unknown x. We refer to approximating distributions that do not depend on x as *admissible*, similar to valid estimators in estimation. From our earlier discussion, we know that information divergence $D(p_{\mathsf{y}}(\cdot;x)\|q(\cdot))$ serves as a natural measure of "goodness" of $q(\cdot)$. It effectively quantifies the modeling loss due to using q instead of the correct distribution $p_{\mathsf{y}}(\cdot;x)$. Thus, we seek to minimize the information divergence over all choices of $q(\cdot)$ that do not depend on x. Here we develop a fundamental approach to such modeling problems. Interestingly, with this approach, we neither require nor benefit from explicitly estimating the unknown parameter x for the purposes of inference.

15.1 Modeling via Mixtures

The key to our approach is to model y as a mixture of the candidate models (i.e., the models in the class that contains the true one), i.e.,

$$q_w(y) = \sum_{x\in\mathcal{X}} w(x)\, p_{\mathsf{y}}(y;x),$$

where the weights $w(x)$ are all nonnegative and $\sum_x w(x) = 1$. A Bayesian viewpoint would consider w to be a prior. Indeed, $w \in \mathcal{P}^{\mathcal{X}}$ since by definition $\mathcal{P}^{\mathcal{X}}$ is the set of functions $p\colon \mathcal{X} \mapsto \mathbb{R}$ such that $p(x) \geq 0$ for all $x \in \mathcal{X}$, and $\sum_{x\in\mathcal{X}} p(x) = 1$. Such an interpretation is not needed at this point, but we will return to it later in the analysis.

Such mixtures can have quite desirable properties, as the following example illustrates.

Example 2. Suppose $\mathsf{y}_n \in \{0,1\} \sim \mathcal{B}(x)$, and $x \in \{0,1\}$. Then the two distributions in the class are very far apart. Indeed,

$$D(p_{\mathsf{y}}(\cdot;0)\|p_{\mathsf{y}}(\cdot;1)) = p_{\mathsf{y}}(0;0)\log\frac{p_{\mathsf{y}}(0;0)}{p_{\mathsf{y}}(0;1)} + p_{\mathsf{y}}(1;0)\log\frac{p_{\mathsf{y}}(1;0)}{p_{\mathsf{y}}(1;1)} = \infty, \tag{4}$$

and, similarly $D(p_{\mathsf{y}}(\cdot;1)\|p_{\mathsf{y}}(\cdot;0)) = \infty$. However, the mixture distribution

$$q(y) = \frac{1}{2}p_{\mathsf{y}}(y;0) + \frac{1}{2}p_{\mathsf{y}}(y;1) \tag{5}$$

has the property that it is quite close to both candidate models, i.e.,

$$D(p_{\mathsf{y}}(\cdot;0)\|q(\cdot)) = D(p_{\mathsf{y}}(\cdot;1)\|q(\cdot)) = 1 \text{ bit!} \tag{6}$$

Evidently, this example runs counter to our intuition from the Euclidean geometry.

The following theorem justifies the use of mixtures as an inherently good approach to modeling. In particular, it establishes that no model can perform better than a mixture one for purposes of inference.

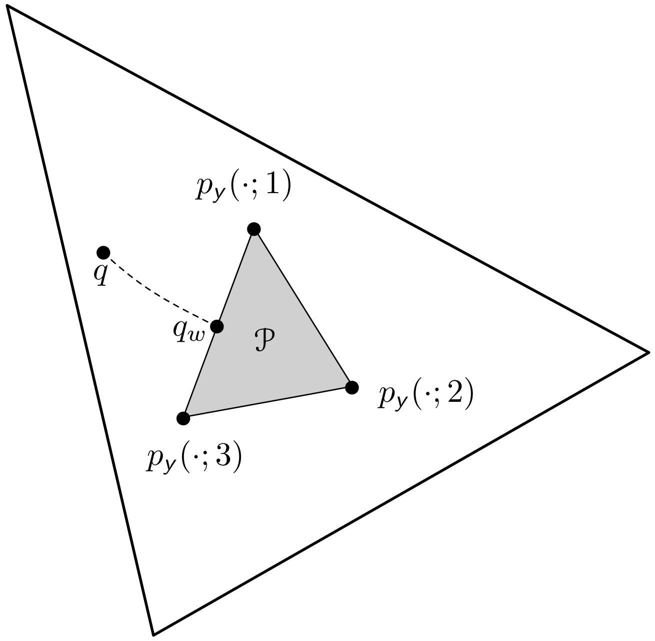

Figure 1: The I-projection of a non-mixture model q onto $\mathcal{P}$ produces a mixture model q_w, which is closer (in a divergence sense) to each of the possible models than q. In this example, $\mathcal{X} = \{1, 2, 3\}$.

Theorem 1. *Let $\{p_y(\cdot; x), x \in \mathcal{X}\}$ be a class of models, and let $q \in \mathcal{P}^{\mathcal{Y}}$ be any admissible distribution (i.e., one that does not depend on the parameter x). Then there exist weights $w(\cdot)$ in a mixture model*

$$q_w(\cdot) = \sum_{x \in \mathcal{X}} w(x)\, p_y(\cdot; x)$$

such that

$$D(p_y(\cdot; x) \| q_w(\cdot)) \leq D(p_y(\cdot; x) \| q(\cdot)) \quad \textit{for all } x \in \mathcal{X}.$$

Proof. Let $\mathcal{P}$ denote the set of distributions that are mixtures of the candidate models $p_y(\cdot; x)$. It is easy to see that $\mathcal{P}$ is a closed and convex set. The theorem is trivially true if $q \in \mathcal{P}$. Let us assume that $q \notin \mathcal{P}$, and define

$$q_w = \arg\min_{p \in \mathcal{P}} D(p \| q), \tag{7}$$

i.e., q_w is the I-projection of q onto $\mathcal{P}$. Clearly, q_w is a mixture distribution. These distributions are depicted in Fig. 1.

But by the Pythagoras' theorem of information geometry,

$$D(p \| q) \geq D(p \| q_w) + D(q_w \| q) \geq D(p \| q_w) \tag{8}$$

for all $p \in \mathcal{P}$. Hence, since $p_y(\cdot; x) \in \mathcal{P}$ for each $x \in \mathcal{X}$, it follows that

$$D(p_y(\cdot;x)\|q(\cdot)) \geq D(p_y(\cdot;x)\|q_w(\cdot)) \quad \text{for all } x \in \mathcal{X}, \tag{9}$$

as Fig. 1 illustrates. □

15.2 Optimization Framework

A natural means for choosing q is according to a minimax criterion, which minimizes the worst-case loss. In particular, we let it be determined by the outcome of a (zero-sum) game between us and nature: we choose q to try to make the divergence as small as possible, knowing that nature will then choose $x \in \mathcal{X}$ to make the divergence as large as possible given our choice. The outcome of this game, viz.,

$$R^+ \triangleq \min_{q \in \mathcal{P}^{\mathcal{Y}}} \left\{ \max_{x \in \mathcal{X}} D(p_y(\cdot;x)\|q(\cdot)) \right\} \tag{10}$$

is sometimes referred to as the *minimax redundancy* in q, since it turns out to correspond, in some sense, to the extra number of bits required to describe the observations when we approximate their distribution by q rather than the true distribution.[1]

Solving for q directly from (10) is difficult. However, conveniently, and as we now develop, the same q is the solution to a different, but closely related, optimization problem that is comparatively easy to solve.

We begin by focusing on the inner maximization in (10), noting that

$$\max_{x \in \mathcal{X}} D(p_y(\cdot;x)\|q(\cdot)) = \max_{w \in \mathcal{P}^{\mathcal{X}}} \sum_x w(x)\, D(p_y(\cdot;x)\|q(\cdot)). \tag{11}$$

Indeed, if x^* is an optimizing x on the left-hand side of (11), an optimizing w^* on the right-hand side of (11) must be

$$w^*(x) = \mathbb{1}_{x=x^*} = \begin{cases} 1 & x = x^*, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

We can therefore rewrite (10) using (11) as

$$R^+ = \min_{q \in \mathcal{P}^{\mathcal{Y}}} \max_{w \in \mathcal{P}^{\mathcal{X}}} \sum_x w(x)\, D(p_y(\cdot;x)\|q(\cdot)). \tag{13}$$

From this form, we can now apply the following saddle-point theorem due to Gallager, which establishes the equivalence of minimax and maximin redundancy, allowing us to interchange the order of minimization and maximization. And as we will see, it is this maximin version of the optimization that is easier to perform directly.

[1] More generally, the terms “approximation loss” and “redundancy” are interchangeable.

Theorem 2 (Redundancy-Capacity Theorem)**.**

$$R^+ = \min_{q\in\mathcal{P}^{\mathcal{Y}}} \max_{w\in\mathcal{P}^{\mathcal{X}}} \sum_{x\in\mathcal{X}} w(x)\, D(p_{\mathsf{y}}(\cdot;x)\|q(\cdot)) \\ = \max_{w\in\mathcal{P}^{\mathcal{X}}} \min_{q\in\mathcal{P}^{\mathcal{Y}}} \sum_{x\in\mathcal{X}} w(x)\, D(p_{\mathsf{y}}(\cdot;x)\|q(\cdot)) \triangleq R^-, \quad (14)$$

and, moreover, the optimizing q and w on both sides of (14) *are the same.*

Recalling the Minimax Inequality—that a minimax quantity always upper bounds the corresponding maximin quantity, the key to the establishing the theorem is to show that the Minimax Inequality holds with equality. Von Neumann (1928) was the first to develop classes of minimax problems in which such equality holds. For example, it holds when the domains are compact, and the objective function is convex in one variable and concave in the other, which is the case our setting. However, rather than establishing Theorem 2 this way, we will follow an approach that leverages useful geometric insights. Accordingly, we first develop these insights.

In particular, let us evaluate the maximin redundancy R^-, focusing first on the inner minimization in its definition. For a fixed set of weights w, we can express the inner objective function on the right-hand side of (14) in the form (14):

$$\begin{aligned}\varphi_{\text{inner}}(q) &\triangleq \sum_x w(x)\, D(p_{\mathsf{y}}(\cdot;x)\|q(\cdot)) \\ &= \sum_{x,y} w(x)\, p_{\mathsf{y}}(y;x) \log \frac{p_{\mathsf{y}}(y;x)}{q(y)} \\ &= \sum_y \sum_x w(x)\, p_{\mathsf{y}}(y;x) \log \frac{p_{\mathsf{y}}(y;x)}{q(y)} \\ &= c - \sum_y q_w(y) \log q(y) && (15) \\ &= c - \mathbb{E}_{q_w}\left[\log q(\mathsf{y})\right], && (16)\end{aligned}$$

where in (15)

$$c \triangleq \sum_y \sum_x w(x)\, p_{\mathsf{y}}(y;x) \log p_{\mathsf{y}}(y;x),$$

which is a constant (i.e., doesn't depend on q), and where

$$q_w(y) \triangleq \sum_x w(x)\, p_{\mathsf{y}}(y;x),$$

which is a mixture distribution with weights w. Applying Gibbs' inequality to (16), we find q^* that minimizes $\varphi_{\text{inner}}(\cdot)$ to be exactly the mixture distribution with weights

w, viz.,

$$q^*(\cdot) = q_w(\cdot) = \sum_x w(x)\, p_{\mathsf{y}}(\cdot; x), \tag{17}$$

so that

$$\min_{q\in\mathcal{P}^{\mathcal{Y}}} \sum_{x\in\mathcal{X}} w(x) D(p_{\mathsf{y}}(\cdot; x)\|q(\cdot)) = \sum_{x\in\mathcal{X}} w(x)\, D(p_{\mathsf{y}}(\cdot; x)\|q_w(\cdot)). \tag{18}$$

Next, we consider the outer maximization in the definition of R^-. We can now express the cost function in the following form that lends itself naturally to optimization:

$$\begin{aligned} \varphi_{\text{outer}}(w) &\triangleq \sum_x w(x)\, D(p_{\mathsf{y}}(\cdot; x)\|q_w(\cdot)) & (19)\\ &= \sum_x \sum_y w(x)\, p_{\mathsf{y}}(y; x) \log \frac{p_{\mathsf{y}}(y; x)}{\sum_{x'} w(x')\, p_{\mathsf{y}}(y; x')}. & (20) \end{aligned}$$

We might attempt to directly determine the optimal weights w^* that maximize $\varphi_{\text{outer}}(w)$, but it turns out that in fact more useful intuition is obtained by returning to a Bayesian view.

15.3 The Least Informative Prior and Model Capacity

Let us interpret w as a prior on x, i.e., $w = p_{\mathsf{x}}$, and let our class of models be expressed as a conditional distribution, i.e., $p_{\mathsf{y}}(y; x) = p_{\mathsf{y}|\mathsf{x}}(y|x)$. This implies that q_w can be interpreted as p_{y}, i.e.,

$$q_w(y) = p_{\mathsf{y}}(y) = \sum_x p_{\mathsf{x}}(x)\, p_{\mathsf{y}|\mathsf{x}}(y|x). \tag{21}$$

From this corresponding Bayesian model we then have

$$\begin{aligned} \varphi_{\text{outer}}(p_x) &= \sum_x p_{\mathsf{x}}(x)\, D(p_{\mathsf{y}|\mathsf{x}}(\cdot|x)\|p_{\mathsf{y}}(\cdot)) \\ &= \sum_{x,y} p_{\mathsf{x}}(x)\, p_{\mathsf{y}|\mathsf{x}}(y|x) \log \frac{p_{\mathsf{y}|\mathsf{x}}(y|x)}{p_{\mathsf{y}}(y)} \\ &= \sum_{x,y} p_{\mathsf{x},\mathsf{y}}(x, y) \log \frac{p_{\mathsf{x},\mathsf{y}}(x, y)}{p_{\mathsf{x}}(x) p_{\mathsf{y}}(y)} \\ &= D(p_{\mathsf{x},\mathsf{y}}(\cdot,\cdot)\|p_{\mathsf{x}}(\cdot)\, p_{\mathsf{y}}(\cdot)) = I(\mathsf{x}; \mathsf{y}), \end{aligned}$$

and thus

$$R^- = \max_{p_{\mathsf{x}}} I(\mathsf{x}; \mathsf{y}), \tag{22}$$

where the optimizing p_{x} is the weight function w^* in the mixture model, and where p_{y} is the corresponding mixture model itself, q^*. Thus, once we establish Theorem 2, i.e., $R^+ = R^-$, we will have established a fundamental link between the cost function we would like to optimize in modeling and the information measures that quantify log-loss due to approximating the true distribution with the optimal mixture.

Because of the central role (22) plays in modeling, the quantities involved are referred to using particular terms.

Definition 1. *Let $p_{\mathsf{y}|\mathsf{x}}$ be a model. The* least informative prior *p_{x}^* for $p_{\mathsf{y}|\mathsf{x}}$ is given by*

$$p_{\mathsf{x}}^* = \arg\max_{p_{\mathsf{x}}} I(\mathsf{x};\mathsf{y}). \tag{23}$$

Definition 2. *The* capacity *C of the model $p_{\mathsf{y}|\mathsf{x}}$ is the (average) cost reduction associated with the least informative prior, i.e.,*

$$C = \max_{p_{\mathsf{x}}} I(\mathsf{x};\mathsf{y}). \tag{24}$$

Some remarks are worthwhile at this point. First, note that C is a measure of the richness of the model. In particular, large values of C correspond to larger classes of models. On the one hand, we want C large so that we are confident the true model is included in it. However, there is a tension—our results also establish that smaller values of C, corresponding to smaller classes of models, allow higher quality inference to be performed (via the optimum approximating distribution).

Additionally, it is worth noting that the model capacity satisfies

$$0 \le C \le \log|\mathcal{X}|, \tag{25}$$

where the upper bound is tight if x can be uniquely determined from any realization $\mathsf{y} = y$ (in which case w^* is a uniform weighting over the alphabet), and where the lower bound is tight if *no* information about x can be inferred from any realization $\mathsf{y} = y$, i.e., $p_{\mathsf{y}}(y;x) = q(y)$ for all x. The upper bound in (25) follows from $I(\mathsf{x};\mathsf{y}) = H(\mathsf{x}) - H(\mathsf{x}|\mathsf{y})$ by noting that the second term is zero and that the uniform distribution maximizes entropy. The lower bound in (25) follows by observing that $H(\mathsf{x}) = H(\mathsf{x}|\mathsf{y})$ in this case, and recalling that mutual information is nonnegative.

It should also be emphasized that even when x is continuous-valued, provided y is discrete the associated least informative prior p_{x} (and hence the optimizing weight distribution w^*) is discrete with at most $|\mathcal{Y}|$ mass points.

In later developments, we will be interested in the case where we have a sequence of N observations $\mathsf{y}_1, \mathsf{y}_2, \ldots, \mathsf{y}_N$, each from some alphabet $\mathcal{Y}$. Note that in such cases both the model capacity C and the optimizing weights (and thus the associated least informative prior) depend on N, which is sometimes referred to as the "horizon" of the problem.

Finally, for $|\mathcal{X}| < \infty$, the optimum model satisfies the following "equidistance" property, which we depict in Fig. 2.

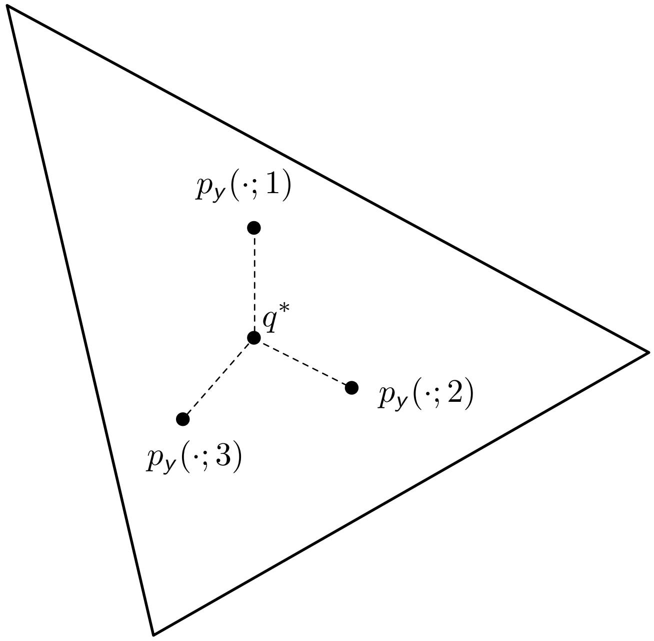

Figure 2: The equidistance property of the optimum mixture model. In this example, $\mathcal{X} = \{1, 2, 3\}$.

Theorem 3 (Equidistance Property)**.** *The optimum mixture model q^*, for which the optimum weights are w^*, is such that*

$$D(p_y(\cdot; x)\|q^*(\cdot)) \leq C \quad \textit{for all } x \in \mathcal{X} \tag{26}$$

with equality for all x such that $w^*(x) > 0$.

Theorem 3 can be interpreted as saying that the approximating distribution does not favor one candidate model over the others. By analogy with simple mechanics, one can interpret the maximization over p_x as corresponding to distributing a unit "mass" (w) over the models in the class such that the "centroid" (p_y) is equidistant from all models (with nonzero mass). Note that zero masses are not required unless $|\mathcal{Y}|$ is too small relative to $|\mathcal{X}|$, which is not the typical scenario of interest.

Proof. Theorem 3 is conceptually straightforward to establish. First, p_x^* (i.e., w^*) is the distribution that maximizes

$$I(x; y) = \sum_{x,y} p_x(x) p_{y|x}(y|x) \log \frac{p_{y|x}(y|x)}{\sum_{x'} p_x(x') p_{y|x}(y|x')} \tag{27}$$

over p_x such that the normalization constraint $\sum_x p_x(x) = 1$ holds, and $p_x(x) \geq 0$ for all $x \in \mathcal{X}$. Using a Lagrange multiplier for the normalization constraint, we have equivalently that p^*_{x} (i.e., w^*) is the distribution that maximizes

$$\varphi(p_{\mathsf{x}}, \lambda) = I(\mathsf{x}; \mathsf{y}) - \lambda \Big[\sum_x p_{\mathsf{x}}(x) - 1 \Big], \tag{28}$$

over p_{x} such that $p_{\mathsf{x}}(x) \geq 0$ for all x, with the constant λ chosen so that the normalization constraint is met.

As shown in Appendix 15.6, $I(\mathsf{x}; \mathsf{y})$, and thus $\varphi(p_{\mathsf{x}}; \lambda)$, is a concave function of p_{x}. Hence, from geometrical considerations, when the maximum of $\varphi(p_{\mathsf{x}}; \lambda)$ with respect to p_{x} occurs at a p^*_{x} away from a particular boundary of the simplex $\mathcal{P}^{\mathcal{X}}$ (i.e., at a p^*_{x} such that $p^*_{\mathsf{x}}(x) > 0$ for some particular x), we must be at a stationary point in that coordinate, i.e.,

$$\left.\frac{\partial \varphi(p_{\mathsf{x}}, \lambda)}{\partial p_{\mathsf{x}}(a)}\right|_{p_{\mathsf{x}} = p^*_{\mathsf{x}}} = \left.\frac{\partial I(\mathsf{x}; \mathsf{y})}{\partial p_{\mathsf{x}}(a)}\right|_{p_{\mathsf{x}} = p^*_{\mathsf{x}}} - \lambda = 0, \quad \text{for all } a \in \mathcal{X} \text{ such that } p^*_{\mathsf{x}}(a) > 0. \tag{29a}$$

However, when the maximum occurs at a p^*_{x} on a boundary of $\mathcal{P}^{\mathcal{X}}$ (i.e., for a p^*_{x} such that $p^*_{\mathsf{x}}(x) = 0$ for some $x \in \mathcal{X}$, then the derivative in that coordinate need only be nonincreasing, i.e.,

$$\left.\frac{\partial I(\mathsf{x}; \mathsf{y})}{\partial p_{\mathsf{x}}(a)}\right|_{p_{\mathsf{x}} = p^*_{\mathsf{x}}} - \lambda \leq 0, \quad \text{for all } a \in \mathcal{X} \text{ such that } p^*_{\mathsf{x}}(a) = 0. \tag{29b}$$

The necessity and sufficiency of (29) are straightforward to formally verify in this case.[2] Proceeding, substituting

$$\begin{aligned}
\frac{\partial I(\mathsf{x}; \mathsf{y})}{\partial p_{\mathsf{x}}(a)} &= \sum_y p_{\mathsf{y}|\mathsf{x}}(y|a) \log \frac{p_{\mathsf{y}|\mathsf{x}}(y|a)}{\sum_{x'} p_{\mathsf{x}}(x')\, p_{\mathsf{y}|\mathsf{x}}(y|x')} \\
&\qquad - \log(e) \sum_y p_{\mathsf{x}}(a)\, p_{\mathsf{y}|\mathsf{x}}(y|a) \cdot \frac{p_{\mathsf{y}|\mathsf{x}}(y|a)}{\sum_{x'} p_{\mathsf{x}}(x')\, p_{\mathsf{y}|\mathsf{x}}(y|x')} \\
&\qquad\quad - \log(e) \sum_{x \neq a} \sum_y p_{\mathsf{x}}(x)\, p_{\mathsf{y}|\mathsf{x}}(y|x) \cdot \frac{p_{\mathsf{y}|\mathsf{x}}(y|a)}{\sum_{x'} p_{\mathsf{x}}(x')\, p_{\mathsf{y}|\mathsf{x}}(y|x')} \\
&= \sum_y p_{\mathsf{y}|\mathsf{x}}(y|a) \log \frac{p_{\mathsf{y}|\mathsf{x}}(y|a)}{\sum_{x'} p_{\mathsf{x}}(x')\, p_{\mathsf{y}|\mathsf{x}}(y|x')} - \log(e) \sum_y p_{\mathsf{y}|\mathsf{x}}(y|a) \\
&= D(p_{\mathsf{y}|\mathsf{x}}(\cdot; a) \| p_{\mathsf{y}}) - \log e
\end{aligned}$$

into (29) and letting $\lambda' = \lambda + \log e$, we obtain[3]

$$D(p_{\mathsf{y}|\mathsf{x}}(\cdot; x) \| p^*_{\mathsf{y}}) \leq \lambda' \quad \text{for all } x, \tag{30}$$

[2] See, e.g., Gallager's text. Or they can be viewed as a simple application of the Karush-Kuhn-Tucker (KKT) conditions for convex optimization with equality and inequality constraints, where that there are two cases follows from complementary slackness.

[3] Note that $p^*_{\mathsf{y}}(y) \triangleq \sum_x p^*_{\mathsf{x}}(x)\, p_{\mathsf{y}|\mathsf{x}}(y|x)$.

with equality for any x such that $p_{\mathsf{x}}^*(x) > 0$.

Finally, it remains only to show that $\lambda' = C$. To see this, note that since in general

$$\sum_x p_{\mathsf{x}}(x)\, D(p_{\mathsf{y}|\mathsf{x}}(\cdot|x) \| p_{\mathsf{y}}(y)) = \sum_{x,y} p_{\mathsf{x}}(x)\, p_{\mathsf{y}|\mathsf{x}}(y|x) \log \frac{p_{\mathsf{y}|\mathsf{x}}(y|x)}{p_{\mathsf{y}}(y)} = I(\mathsf{x};\mathsf{y}),$$

it follows from (30) and the condition for equality that

$$\begin{aligned}\lambda' &= \sum_{\{x\in\mathcal{X}\colon\, p_{\mathsf{x}}^*(x)>0\}} p_{\mathsf{x}}^*(x)\, D(p_{\mathsf{y}|\mathsf{x}}(\cdot|x) \| p_{\mathsf{y}}^*(y)) \\ &= \sum_{x\in\mathcal{X}} p_{\mathsf{x}}^*(x)\, D(p_{\mathsf{y}|\mathsf{x}}(\cdot|x) \| p_{\mathsf{y}}^*(y)) \\ &= I(\mathsf{x};\mathsf{y})\Big|_{p_{\mathsf{x}}=p_{\mathsf{x}}^*} \\ &= C.\end{aligned}$$

□

Now having the equidistance property, the Redundancy-Capacity Theorem follows almost immediately.

Proof of Theorem 2. Let us denote the optimizing q in the maximin optimization R^- by q_*^- (this is our p_{y}^* in the Bayesian interpretation above). Then by the Equidistance Property, we have

$$\max_{w\in\mathcal{P}^{\mathcal{X}}} \sum_x w(x)\, D(p_{\mathsf{y}}(\cdot;x) \| q_*^-) \le \max_{w\in\mathcal{P}^{\mathcal{X}}} \sum_x w(x)\, C = C,$$

from which we obtain

$$R^+ = \min_{q\in\mathcal{P}^{\mathcal{Y}}} \left[\max_{w\in\mathcal{P}^{\mathcal{X}}} \sum_x w(x)\, D(p_{\mathsf{y}}(\cdot;x) \| q) \right] \le \max_{w\in\mathcal{P}^{\mathcal{X}}} \sum_x w(x)\, D(p_{\mathsf{y}}(\cdot;x) \| q_*^-) \le C = R^-, \tag{31}$$

where the last equality follows from our evaluation of the maximin optimization in Section 15.3. But, from the Minimax Inequality, $R^+ \ge R^-$, so the inequalities in (31) must hold with equality. Hence, $R^+ = R^-$. Moreover, from the first of the inequalities in (31) holding with equality we obtain that q_-^* must also be an optimizing q in the minimax optimization R^+. Likewise, from the second of the inequalities in (31) holding with equality, together with the conditions for equality in the Equidistance Property, we obtain that the optimizing weights w_-^* in the maximin optimization R^- must also be an optimizing w in the minimax optimization R^+. □

15.4 Inference with Mixture Models

As an illustration of how we apply our optimized mixture models for inference, consider a fairly general scenario in which there are a set of variables y_- we observe, and a set y_+ that we do not observe but about which we want to make inferences. In particular, we want to find some $q_{\mathsf{y}_+|\mathsf{y}_-}(\cdot|y_-)$ that is as close as possible in the minimax redundancy sense to

$$p_{\mathsf{y}_+|\mathsf{y}_-}(\cdot|y_-;x) = \frac{p_{\mathbf{y}}(\cdot, y_-; x)}{\sum_b p_{\mathsf{y}}(b, y_-; x)} \tag{32}$$

for all values x of the unknown parameter, where $\mathbf{y} = (\mathsf{y}_+, \mathsf{y}_-)$.

The desired distribution is readily obtained from our general optimized mixture distribution

$$q^*(\mathbf{y}) = \sum_x w^*(x)\, p_{\mathbf{y}}(\mathbf{y}; x). \tag{33}$$

Specifically, we have

$$\begin{aligned}
q_{\mathsf{y}_+|\mathsf{y}_-}(\cdot|y_-) &\triangleq \frac{q_{\mathsf{y}}(y_+, y_-)}{q_{\mathsf{y}_-}(y_-)} && (34)\\
&= \frac{\sum_x w^*(x)\, p_{\mathsf{y}}(y_+, y_-; x)}{\sum_a w^*(a)\, p_{\mathsf{y}_-}(y_-; a)} && (35)\\
&= \sum_x \left[\frac{w^*(x)\, p_{\mathsf{y}_-}(y_-; x)}{\sum_a w^*(a)\, p_{\mathsf{y}_-}(y_-; a)} \right] p_{\mathsf{y}_+|\mathsf{y}_-}(y_+|y_-; x) && (36)\\
&= \sum_x w^*(x|y_-)\, p_{\mathsf{y}_+|\mathsf{y}_-}(y_+|y_-; x) && (37)
\end{aligned}$$

where we have defined the "revised" weights

$$w^*(x|y_-) \triangleq \frac{w^*(x)\, L_{y_-}(x)}{\sum_a w^*(a)\, L_{y_-}(a)}, \tag{38}$$

with

$$L_{y_-}(x) = p_{\mathsf{y}_-}(y_-; x) \tag{39}$$

denoting the likelihood function of the observed data.

Interestingly, we can interpret (38) as a "weighting-by-relative-likelihood" procedure, which is rather different than the naïve maximum-likelihood approach discussed earlier. Indeed in our mixture approach we form our inference as a mixture of inferences from the different candidate models, each weighted according to the relative likelihood of the candidate model based on the observed data. This represents a kind of "soft-decision" modeling. By contrast, the maximum likelihood approach corresponds to the choice

$$w^*(x|y_-) = \begin{cases} 1 & x = \hat{x}_{\mathrm{ML}}(y_-) \\ 0 & \text{otherwise}, \end{cases} \tag{40}$$

where

$$\hat{x}_{\mathrm{ML}}(y_-) = \arg\max_a L_{y_-}(a), \tag{41}$$

which can be interpreted as a kind of "hard-decision" modeling.

15.5 Alternative Weightings

It turns out that in many cases, modeling performance is not too sensitive to the exact details of w^*. As a result, in practice, the choice of weighting is often influenced by implementation and other considerations. In fact, often it is sufficient just to choose a weighting w that includes all models, i.e., $w > 0$, an example of which is the uniform weighting.

Example 3. We now return to Example 1, where $\mathsf{y}_i \sim \mathcal{B}(x)$ are i.i.d. with $\mathcal{Y} = \{0, 1\}$, $\mathcal{X} = [0, 1]$, and $\mathbb{P}(\mathsf{y}_i = 1) = x$. If we use a uniform weighting (prior), it is straightforward to verify that the resulting prediction is of the form

$$q_{\mathsf{y}_n|\mathsf{y}_{n-1},\ldots,\mathsf{y}_1}(1|y_{n-1},\ldots,y_1) = \frac{\int_0^1 x^{n_1+1}(1-x)^{n_0}\,\mathrm{d}x}{\int_0^1 x^{n_1}(1-x)^{n_0}\,\mathrm{d}x} = \frac{n_1+1}{n+1}, \tag{42}$$

where n_0 and n_1 are the number of zeros and ones, respectively, in the observed sequence $y_1, y_2, \ldots, y_{n-1}$.

Note that since (42) can be interpreted as an estimate of x, since x is the probability of a one, our predictor effectively produces an estimate

$$\hat{x}^{(n-1)} = \frac{n_1+1}{n+1}, \tag{43}$$

which corresponds to an estimator for such problems first proposed by Laplace.[4] By contrast, a maximum likelihood approach produces the estimate

$$\hat{x}_{\mathrm{ML}}^{(n-1)} = \frac{n_1}{n-1}, \tag{44}$$

which is less appealing for small values of n. Indeed, in the absence of any data, (43) is clearly preferable.

We should also emphasize that in this particular example, the mixture modeling approach happens to degenerate to an instance of estimate-and-substitute heuristic discussed at the outset of our modeling discussion. However, it is worth emphasizing that this heuristic is, in general, not optimal.

The uniform weighting chosen in Example 3 corresponds to an instance of what viewed as a "maximally ignorant prior." As we will develop more generally, such priors can be developed for different scenarios using information geometry.

[4] In the 18th century, Laplace posed the question "What is the probability that the sun will rise tomorrow?" and developed this estimator as an approach to answering it.

15.6 Appendix: Concavity of Entropy and Mutual Information

First, it is straightforward to establish that $H(p)$ is concave in p using the log-sum inequality. In particular, let p_1 and p_2 be any two distributions on a common alphabet $\mathcal{Y}$, and $p_\lambda = \lambda p_1 + (1-\lambda)p_2$. Then

$$\begin{aligned}
-H(p_\lambda) &= \sum_y p_\lambda(y) \log p_\lambda(y) \\
&= \sum_y [\lambda p_1(y) + (1-\lambda)p_2(y)] \log[\lambda p_1(y) + (1-\lambda)p_2(y)] \\
&\le -\sum_y \left[\lambda p_1(y) \log \frac{\lambda p_1(y)}{\lambda} + (1-\lambda)p_2(y) \frac{(1-\lambda)p_2(y)}{(1-\lambda)}\right] && (45) \\
&= \lambda \sum_y p_1(y) \log p_1(y) + (1-\lambda) \sum_y p_2(y) \log p_2(y) \\
&= -\left[\lambda H(p_1) + (1-\lambda)H(p_2)\right], && (46)
\end{aligned}$$

where to obtain (45) we have used the log-sum inequality with $K = 2$, $u_1 = \lambda p_1(y)$, $u_2 = \lambda p_2(y)$, $v_1 = \lambda$, and $v_2 = 1 - \lambda$. Hence, rewriting (46), we obtain

$$H(\lambda p_1 + (1-\lambda)p_2) \ge \lambda H(p_1) + (1-\lambda)H(p_2),$$

as desired.

Next, mutual information $I(\mathsf{x};\mathsf{y})$ is a function of the joint distribution $p_{\mathsf{x},\mathsf{y}}$, which in turn is a function of p_{x} and $p_{\mathsf{x}|\mathsf{y}}$. Accordingly, we can make this dependence explicit with the alternative notation $I(p_{\mathsf{x}}, p_{\mathsf{y}|\mathsf{x}})$, i.e.,

$$I(p_{\mathsf{x}}, p_{\mathsf{y}|\mathsf{x}}) = D(p_{\mathsf{x},\mathsf{y}} \| p_{\mathsf{x}}\, p_{\mathsf{y}}) = \sum_{x,y} p_{\mathsf{x}}(x)\, p_{\mathsf{y}|\mathsf{x}}(y|x) \log \frac{p_{\mathsf{y}|\mathsf{x}}(y|x)}{\sum_a p_{\mathsf{x}}(a)\, p_{\mathsf{y}|\mathsf{x}}(y|a)}. \qquad (47)$$

We now show that $I(p_{\mathsf{x}}, p_{\mathsf{y}|\mathsf{x}})$ is concave in p_{x}. To see, this, note that in our equivalent notation we can express the expansion $I(\mathsf{x};\mathsf{y}) = H(\mathsf{y}) - H(\mathsf{y}|\mathsf{x})$ in the form

$$I(p_{\mathsf{x}}, p_{\mathsf{y}|\mathsf{x}}) = H\left(\sum_x p_{\mathsf{x}}(x)\, p_{\mathsf{y}|\mathsf{x}}(y|x)\right) - \sum_x p_{\mathsf{x}}(x)\, H(p_{\mathsf{y}|\mathsf{x}}(\cdot|x)). \qquad (48)$$

But the argument of $H(\cdot)$ in the first term of (48) is a linear function of p_{x}, and thus the first term is concave since entropy is concave. Moreover, the second term in (48) is linear in p_{x}. Thus, the sum of the two terms is concave in p_{x}, as desired.

15.7 Further reading

For further detail on the modeling, the tutorial paper by Feder and Merhav is useful, if somewhat advanced, reading. The original proof of the redundancy-capacity theorem appears in the unpublished 1979 manuscript: R. G. Gallager, "Source Coding with Side Information and Universal Coding," (Tech. Rep. LIDS-P-937). The second edition of T. Cover and J. Thomas, *Elements of Information Theory* also has a development of the redundancy-capacity theorem. For further discussion of the equidistance property of the optimum mixture model, and related insights, see the book R. G. Gallager, *Information Theory and Reliable Communication*, 1968.