6.780 Notes

Lecturer: Greg Wornell

Andrew Liu

Spring 2024

My notes for 6.780, "Inference and Information". The instructor for this course was Gregory Wornell (https://web.mit.edu/gww/www/). All credit for these notes goes to the teaching staff!

Last updated on Friday 5th April, 2024.

Contents

1 February 6, 2024

2 February 8, 2024

The setup for today's lecture is a model family $\mathcal{H} \in \{H_0, H_1, \ldots, H_{M-1}\}$. In the classification problem, we can think of $\mathcal{H}$ as a set of class labels, and we want to determine the correct label given some test data.

2.1 Bayesian Binary Hypothesis Testing

In this case, $M = 2$, so there are only two hypotheses. Our model has two major components. The first is some a priori information

$$\begin{aligned} P_0 &= \mathbb{P}[H = H_0] \\ P_1 &= \mathbb{P}[H = H_1] = 1 - P_0. \end{aligned}$$

We also have the observation model, which is given by likelihood functions

$$\begin{aligned} H_0 &: p_{Y|H}(\cdot|H_0) \\ H_1 &: p_{Y|H}(\cdot|H_1). \end{aligned}$$

Our goal is to create a **decision rule**, a.k.a. a **classifier**, which maps every $y \in \mathcal{Y}$ to some hypothesis $H_i \in \mathcal{H} = \{H_0, H_1\}$. This is somewhat confusing with the standard terminology of a hypothesis class being the set of possible solutions to a model, but we accept it for now.

Definition 2.1 (Cost)
In its most general form, we let

$$C(H_j, H_i) \triangleq C_{ij}$$

denote the cost of predicting H_j when the correct class is H_i.

Using cost to drive the notion of "best", our best possible decision rule takes

the form

$$\hat{H}(\cdot) = \arg\min_f \mathbb{E}_{Y,H}[C(H, f(Y))].$$

The expected cost on the RHS is called **Bayes risk**, which we denote as $\varphi(f)$ for any decision rule f.

We can explicitly calculate this quantity:

$$\begin{aligned}\varphi(f) &= \mathbb{E}_{Y,H}[C(H, f(Y))] \\ &= \mathbb{E}_Y[\mathbb{E}_{H|Y}[C(H, f(Y))|Y = y]] \\ &= \int p_Y(y)\mathbb{E}[C(H, f(Y))|Y = y]\mathrm{d}y.\end{aligned}$$

Notice that we have control over the expected risk for each point, so to minimize $\varphi(f)$, we only have to solve a solution for individual points. For a fixed $y* \in \mathcal{Y}$, there are two possibilities; if $f(y^*) = H_0$, then

$$\mathbb{E}[C(H, f(y^*))|y = y^*] = C_{00}\mathbb{P}[H = H_0|y = y^*] + C_{01}\mathbb{P}[H = H_1|y = y^*],$$

otherwise

$$\mathbb{E}[C(H, f(y^*))|y = y^*] = C_{10}\mathbb{P}[H = H_0|y = y^*] + C_{11}\mathbb{P}[H = H_1|y = y^*].$$

This already technically gives us the optimal decision rule; for any given input y, we can explicitly compute both values, and return the hypothesis that gives the lesser of the two values. We can also express this in a simpler form. Since

$$\mathbb{P}[H = H_i|Y = y] = \frac{p_{Y|H}(y|H_i)p_H(H_i)}{p_Y(y)},$$

we can substitute into the above expressions:

$$C_{00}p_{Y|H}(y|H_0)P_0 + C_{01}p_{Y|H}(y|H_1)P_1 \underset{\hat{H}=H_0}{\overset{\hat{H}=H_1}{\gtrless}} C_{10}p_{Y|H}(y|H_0)P_0 + C_{11}p_{Y|H}(y|H_1)P_1$$

We can rewrite this expression in terms of the ratios

$$L(y) \triangleq \frac{p_{Y|H}(y|H_1)}{p_{Y|H}(y|H_0)} \underset{\hat{H}=H_0}{\overset{\hat{H}=H_1}{\gtrless}} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} \triangleq \eta.$$

We call $L(y)$ the **likelihood ratio**.

Theorem 2.2 (Likelihood Ratio Test)
Given a priori probabilities P_0, P_1, data y, observation models $p_{Y|H}(\cdot|H_0), p_{Y|H}(\cdot|H_1)$, and costs $C_{00}, C_{01}, C_{10}, C_{11}$, the Bayesian decision rule form

$$L(y) \triangleq \frac{p_{Y|H}(y|H_1)}{p_{Y|H}(y|H_0)} \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} \triangleq \eta,$$

meaning that the decision is $\hat{H}(y) = H_1$ when $L(y) > \eta$, $\hat{H}(y) = H_0$ when $L(y) < \eta$, and it is indifferent when $L(y) = \eta$.

Note that the optimal rule is simple and deterministic. Prof. Wornell makes a point about $L(y)$ being a scalar that we can always calculate. This is the heart of classification models; in larger neural nets, like ImageNet, ultimately what the large network of weights allows us to do is to express the intractable probabilities and compute a scalar value.

2.2 0-1 Loss

In the case of "0-1 loss", i.e., $C_{00} = C_{11} = 0$, $C_{01} = C_{10} = 1$, in which case our test simplifies to

$$p_{H|Y}(H_1|y) \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} p_{H|Y}(H_0|y).$$

This is the **maximum a posteriori** (MAP) decision rule.

If we additionally assume that $P_0 = P_1$, i.e., that our prior belief is indifferent, then our test further simplifies to

$$p_{Y|H}(y|H_1) \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} p_{Y|H}(y|H_0).$$

This is the **maximum likelihood** (MLE) decision rule. In either case, the expected rate of error is given by

$$\varphi(\hat{H}) = \mathbb{P}[\hat{H}(Y) = H_0, H = H_1] + \mathbb{P}[\hat{H}(Y) = H_1, H = H_0].$$

Example 2.3 (Communicating a Bit)
We have a signal y, randomly distributed with variance σ^2, and with two possible sources s_0, s_1.

The likelihood ratio test gives

$$\ln L(y) = \ln\left(\frac{e^{-(y-s_1)^2/(2\sigma^2)}}{e^{-(y-s_0)^2/(2\sigma^2)}}\right) = \frac{1}{2\sigma^2}((y-s_0)^2 - (y-s_1)^2).$$

Assuming $0-1$ loss, $\ln L(y) = 0$, so the decision boundary is

$$y \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \frac{s_0+s_1}{2}.$$

We could compute the expected rate of error as follows:

$$\begin{aligned}
\varphi(\hat{H}) &= \frac{1}{2}\left(\mathbb{P}[\hat{H}(Y) = H_0 | H = H_1] + \mathbb{P}[\hat{H}(Y) = H_1 | H = H_0]\right) \\
&= \frac{1}{2}\left(\mathbb{P}\left[y < \frac{s_0+s_1}{2}\middle| H = H_1\right] + \mathbb{P}\left[y \geq \frac{s_0+s_1}{2}\middle| H = H_0\right]\right) \\
&= \frac{1}{2}\left(\mathbb{P}\left[\frac{y-s_1}{\sigma} < \frac{s_0-s_1}{2\sigma}\middle| H = H_1\right] + \mathbb{P}\left[\frac{y-s_0}{\sigma} \geq \frac{s_1-s_0}{2\sigma}\middle| H = H_0\right]\right) \\
&= Q\left(\frac{s_1-s_0}{2\sigma}\right).
\end{aligned}$$

The quantity $(s_1 - s_0)/\sigma$ is a measure of signal-to-noise; the larger the SNR, the more uncertain we are about our prediction, i.e., the higher our expected rate of error.

3 February 13, 2023

Bad weather day today. Prof. Wornell tells us about how this is the first time he's given a lecture over zoom since the pandemic. He was hoping that he would never have had to give a zoom lecture again, but here we are.

3.1 Review: Bayesian Hypothesis Testing

Overarching goal: optimize the expected cost of choosing one hypothesis over another. This can be accomplished with the Likelihood Ratio Test:

$$L(y) \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \eta$$

Some issues with this solution: we require some notion of costs, as well as priors $P_i = P_H(H_i)$. It can be difficult to assign probabilities to abstract concepts, like $\mathbb{P}$[patient has disease].

3.2 Non-Bayesian Hypothesis Testing

The "folk theorem" that we will be proving today is that all optimum decision rules takes the form of a Likelihood Ratio Test:

$$\frac{p_{Y|H}(y|H_1)}{p_{Y|H}(y_{H_0})} \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \eta,$$

for some η. This is largely true, but not always.

3.3 General Performance Measures

Let $\hat{H}(\cdot)$ be any rule. An equivalent characterization of this rule is some partition of the observation space Y:

$$y_0 = \{y \in Y : \hat{H}(y) = H_0\}$$
$$y_1 = y \backslash y_0$$

Then $P_D = \mathbb{P}[\hat{H} = H_1 | H = H_1] = \int_{y_1} p_{Y|H}(y|H_1)\mathrm{d}y$ is called the **detection probability**, and $P_F = \mathbb{P}[\hat{H} = H_1 | H = H_0] = \int_{y_1} p_{Y|H}(y|H_0)\mathrm{d}y$ is called the **"false alarm" probability**. Some related terminology: $P_M = 1 - P_D$ is called the "miss" probability.

- Statistics terminology: $P_E^1 = P_F$, $P_E^2 = P_M$, "probability of error of each kind". The probability of error of the first kind is called the "size" of the test, while the probability of error the second kind is called the "power" of the test.
- Medical terminology: P_F is the false positive rate, while P_M is the false negative rate.
- Learning / pattern classification: $P_R = P_D$ is the "recall" or "sensitivity". The "precision" is defined as $P_P = \mathbb{P}[H = H_1 | \hat{H} = H_1] = 1/(1 + P_F/P_D \cdot P_0/P_1)$.

In general, P_D and P_F are conflicting objectives. We seek large P_D and small P_F. The bayesian approach to this "multi-objective" optimization is to choose the rule that satisfies:

$$\min_{\hat{H}(\cdot)}(\alpha P_F - \beta P_D).$$

3.4 Operating Characteristic of the LRT

Other tradeoffs are possible. Consider the family of ratio tests:

$$\{\hat{H}(\cdot) = \text{LRT}, \text{for some } \eta\}.$$

Example 3.1

Consider two hypotheses

$$H_0 : y \sim \mathcal{N}(0, \sigma^2)$$
$$H_1 : y \sim \mathcal{N}(m, \sigma^2)$$

The LRT is given by

$$y \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \frac{m}{2} + \frac{\sigma^2 \ln \eta}{m} \triangleq \gamma.$$

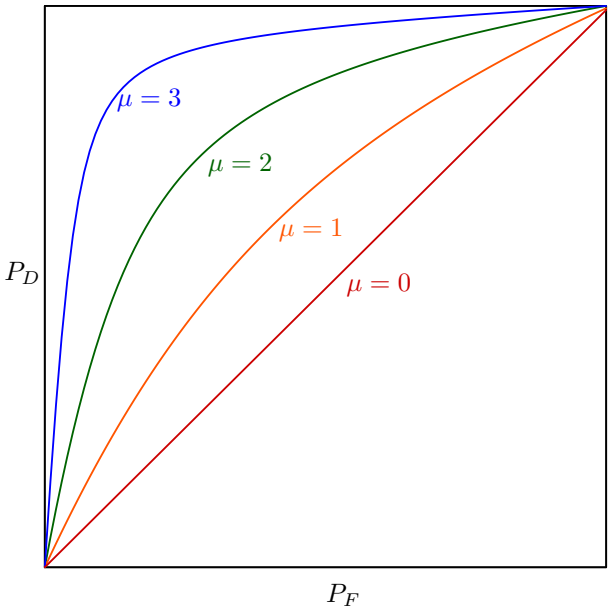
The false positive rate

$$P_F = \int_\gamma^\infty p_{Y|H}(y|H_0)\mathrm{d}y = Q\left(\frac{\gamma}{\sigma}\right),$$

and the detection probability

$$P_D = \int_\gamma^\infty p_{Y|H}(y|H_1)\mathrm{d}y = Q\left(\frac{\gamma - m}{\sigma}\right).$$

Graphing a plot of (P_F, P_D) over different values of $\mu = m/\sigma$ gives



This curve is called the **OC-LRT**: Operating Characteristic of the Likelihood Ratio Test.

$$\text{OC-LRT} : \{(P_F, P_D) : \hat{H}(\cdot) \text{ is LRT for some } \eta\}.$$

Claim 3.2
OC-LRT is monotonic and non-decreasing.

Proof. If $\eta_2 > \eta_1$, then $P_D(\eta_2) \leq P_D(\eta_1)$ and $P_F(\eta_2) \leq P_F(\eta_1)$. □

A few other key properties here:

- All of the frontiers are concave-down, i.e., we should always be "better" at maximizing P_D than minimizing P_F
- As $\mu \to \infty$, we approach the optimal curve, which is when P_D is 1 everywhere and P_F is 0 everywhere. This makes sense intuitively, since μ large separates the two hypotheses.
- Similarly, when $\mu = 0$, the two hypotheses are indistinguishable, so the frontier is as good as random guessing.

- When $P_F = P_D = 1$ when $\gamma = 0$ and $P_F = P_D = 0$ when $\gamma \to \infty$. Therefore, as γ increases to infinity, we travel from top-right to bottom-left along each of the curves (this will always be the case for any OC-LRT, for this reason).

We will prove some of these next lecture.

3.5 Neyman-Pearson Criterion

To avoid the problem of costs and priors, a common alternate criteria choose a rule subject is the **Neyman-Pearson Criterion**:

$$\max_{\hat{H}(\cdot)} P_D \text{ s.t. } P_F \leq \alpha.$$

In words, choose the hypothesis with largest detection power given a fixed upper bound on the false alarm size.

Theorem 3.3 (Neyman-Pearson Lemma, Specialized)
For deterministic $\hat{H}(\cdot)$, a solution to the Neyman-Pearson Criterion is an LRT when the LRT is continuous. In other words,

$$\hat{H}(y) = H_{\mathbb{1}_{L(y)\geq\eta}},$$

where η is the smallest threshold s.t.

$$P_F = \mathbb{P}(L(y) \geq \eta | H = H_0) \leq \alpha.$$

This statement of NP is considered 'specialized' because we are not employing randomization in our hypotheses. We will see and prove the full version next lecture.

Proof. We can prove this with lagrange multipliers. Fix $P_F = \alpha' \leq \alpha$. Then, we want

to optimize

$$\begin{aligned} min_{\hat{H}(\cdot)}\varphi(\hat{H}) &= (1-P_D)+\lambda(P_F-\alpha') \\ &= \int_{y_0} p_{Y|H}(y|H_1)\mathrm{d}y + \lambda\left(\int_{y_1} p_{Y|H}(y|H_0)\mathrm{d}y - \alpha'\right) \\ &= \lambda(1-\alpha') + \int_{y_0}\Big(p_{Y|H}(y|H_1) - \lambda p_{Y|H}(y|H_0)\Big)\mathrm{d}y. \end{aligned}$$

The min φ occurs when we assign y to y_0 whenever the the integrand is ≤ 0, as this minimizes the cost. For the same reason, we want α' to be as large as possible. Therefore,

$$\frac{p_{Y|H}(y|H_1)}{p_{Y|H}(y|H_0)} \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \lambda,$$

where α' is chosen to be the largest threshold achievable by LRT. □

We also present an alternate proof below.

Proof. We compare the LRT decision region with an arbitrary decision region, and show that we cannot do better than the LRT decision region. Let $\mathcal{Y}_1^\eta$ be the set of points for which $\hat{H}^\eta(y) = H_1$, and $\mathcal{Y}_1$ be the set of points for which $\hat{H}(y) = H_1$ for an arbitrary decision rule. We have

$$(\mathbb{1}(L(y)\geq\eta) - \mathbb{1}(y\in\mathcal{Y}_1))(L(y)-\eta)\geq 0,$$

which can be verified with casework. Then,

$$\begin{aligned} &\int_{\mathcal{Y}} (\mathbb{1}(L(y)\geq\eta) - \mathbb{1}(y\in\mathcal{Y}_1))(L(y)-\eta)p_{Y|H}(y|H_0)\mathrm{d}y \\ &= \int_{\mathcal{Y}} (\mathbb{1}(L(y)\geq\eta) - \mathbb{1}(y\in\mathcal{Y}_1))(p_{Y|H}(y|H_1) - \eta p_{Y|H}(y|H_0))\mathrm{d}y \geq 0. \end{aligned}$$

Some expansion and rearranging gives

$$\int_{\mathcal{Y}_1^\eta} p_{Y|H}(y|H_1)\mathrm{d}y - \int_{\mathcal{Y}_1} p_{Y|H}(y|H_1)\mathrm{d}y \geq \eta\left(\int_{\mathcal{Y}_1^\eta} p_{Y|H}(y|H_0)\mathrm{d}y - \int_{\mathcal{Y}_1} p_{Y|H}(y|H_0)\mathrm{d}y\right),$$

which (term-by-term) is just

$$P_D^\eta - P_D \geq \eta(P_F^\eta - P_F).$$

When P_F^η is fixed at α, the constraints of the problem force us to pick $P_F \leq P_F^\eta$, so $P_D^\eta - P_D \geq 0$. But this implies that any hypothesis that is not the LRT can only have worse detection power, hence $\hat{H}_1^\eta$ is a valid solution. □

3.6 Hypothesis Testing in the LRT framework

The typical LRT criteria:

$$L(y) \underset{\hat{H}=H_0}{\overset{\hat{H}=H_1}{\gtrless}} \eta.$$

When we apply a monotonic function to both sides, the criteria remains the same. In particular, when we apply P_F, we get

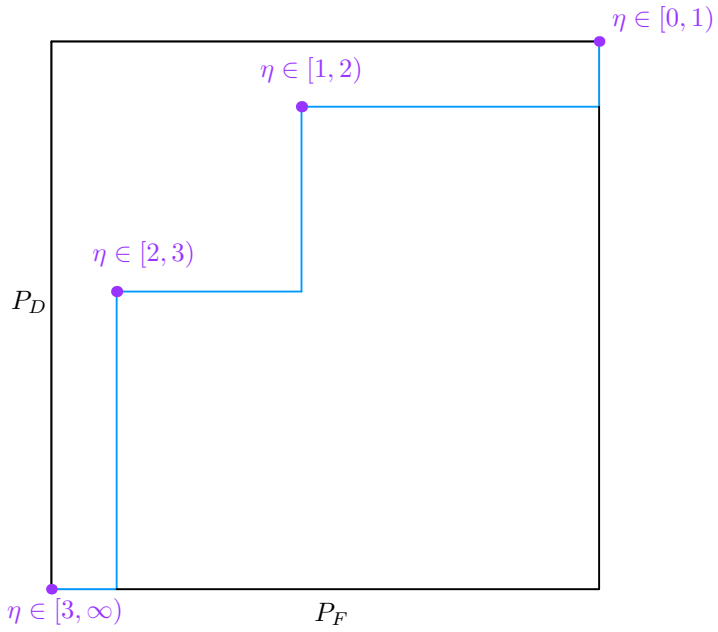
$$p_*(y) \overset{\Delta}{=} P_F(L(y)) \underset{\hat{H}=H_1}{\overset{\hat{H}=H_0}{\gtrless}} P_F(\eta) = \alpha.$$

Note that the direction of the hypothesis changed, since P_F is monotonically decreasing in η. The RHS is the significance level of the test, while p_* is a function that maps each data point to a p-value. Roughly speaking, if p_* is large, this means L was small, so the data was not very significant. If p_* is small, this means L was large, so the data was significant. The threshold of "significant" is determined by our threshold α.

4 February 15, 2024

4.1 Randomizing Neyman-Pearson

First, some intuition on why randomizing can be beneficial. Imagine any scenario involving a discrete process such as a Poisson process. In this case, the points on the OC-LRT are discontinuous, such as in the following diagram:



We emphasize two things here:

- Even though the points on the OC-LRT, represented in purple, are discontinuous, each still corresponds to a continuous range of values η. In this scenario, $P_D = P_F = 0$ when $\eta \geq 3$.
- The Neyman-Pearson function $\zeta_{NP}(\alpha)$ is continuous, which is represented by the blue curve. Specifically, it represents the maximal P_D given that $P_F \leq \alpha$.

Now randomize in the following way. For some $p \in (0,1)$, sample $u \sim \text{Unif}[0,1]$, and choose

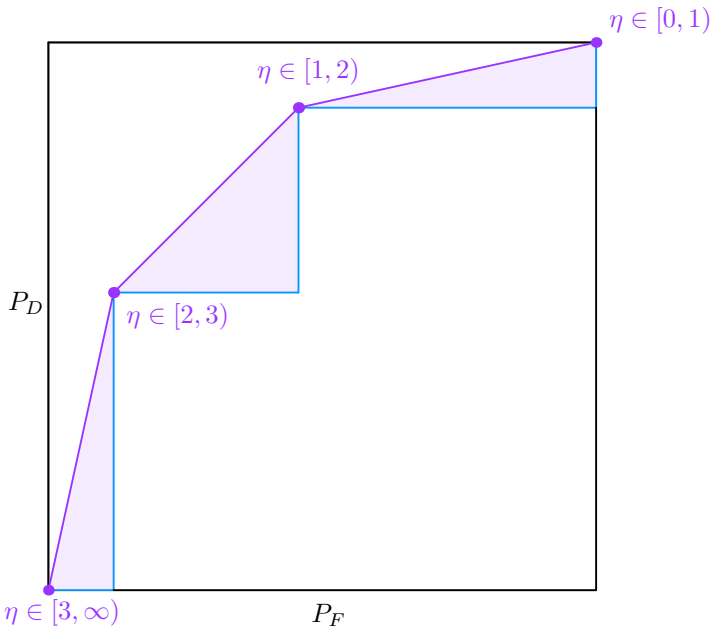
$$\hat{H}(y) = \begin{cases} \hat{H}'(y) & u \leq p \\ \hat{H}''(y) & u > p, \end{cases}$$

where $\hat{H}(y)$ corresponds to the LRT with ratio η' and $\hat{H}''(y)$ corresponds to the LRT with ratio η''. This test achieves

$$P_D = pP_D(\eta') + (1-p)P_D(\eta'')$$
$$P_F = pP_F(\eta') + (1-p)P_F(\eta''),$$

which lies on the line segment between $(P_D(\eta'), P_F(\eta'))$ and $(P_D(\eta''), P_F(\eta''))$. If we apply this randomization for any η_1, η_2 where the points on the OC-LRT becomes

discrete, then we can fill in the "gaps" in the Neyman-Pearson function as shown below



4.2 Full Neyman-Pearson Lemma

Now we are ready for the full version of the Neyman-Pearson Lemma.

Theorem 4.1 (Neyman-Pearson Lemma)
Given hypotheses $p_{Y|H}(y|H_0)$,$p_{Y|H}(y|H_1)$ and $\alpha \in [0,1]$, for some η and $p \in [0,1]$ there exists rule of the form

$$q(y) = \begin{cases} 0 & L(y) < \eta \\ p & L(y) = \eta \\ 1 & L(y) > \eta \end{cases},$$

where $P_F(q_*) = \alpha$ and $P_D(q_*) \geq P_D(q)$ for any decision rule q satisfying $P_F(q) \leq \alpha$.

To see how this is the same setup as our previous example, consider the randomized OC-LRT from the previous example. We have two hypotheses $\hat{H}'(y) = H_0, \hat{H}''(y) = H_1$ which are LRTs for $\eta_1 < \eta_2$ corresponding to distinct, discrete points

on the OC-LRT, e.g., $\eta_i = i$ for $i \in [2]$. To achieve larger P_D for some "middling" restriction on P_F, e.g., $P_F \le \alpha = 1.5$, we need to randomize, and in particular we showed that the randomized hypothesis given by

$$q(y) = \begin{cases} H_0 & L(y) \le \eta_1 \\ p & \eta_1 < L(y) \le \eta_2 \\ H_1 & L(y) > \eta_2, \end{cases}$$

gives us the more "optimal" bound that we want. In the limit where $\eta_1 = \eta_2$, we get the hypothesis in the Neyman-Pearson lemma.

Proof. Assume such a q_* exists. Then we have

$$(q_*(y)-q(y))(p_{Y|H}(y|H_1)-\eta p_{Y|H}(y|H_0)) = (q_*(y)-q(y))(L(y)-\eta)p_{Y|H}(y|H_0) \ge 0 \quad \forall y \in \mathcal{Y}.$$

This can be justified with casework:

- If $L(y) < \eta$, $q_*(y) = 0$, and since $q(y) \ge 0$, we get the product of two non-positive things.
- If $L(y) = \eta$, the product is equal to 0.
- If $L(y) > \eta$, $q_*(y) = 1$, and since $q(y) \le 1$, we get the product of two non-negative things.

Thus

$$\int (q_*(y) - q(y))(p_{Y|H}(y|H_1) - \eta p_{Y|H}(y|H_0))\mathrm{d}y \ge 0,$$

which is the same as

$$\int q_*(y)p_{Y|H}(y|H_1)\mathrm{d}y - \int q(y)p_{Y|H}(y|H_1)\mathrm{d}y \\ \ge \eta\left(\int q_*(y)p_{Y|H}(y|H_0)\mathrm{d}y - \int q(y)p_{Y|H}(y|H_0)\mathrm{d}y\right),$$

i.e.,

$$P_D(q_*) - P_D(q) \ge \eta(P_F(q_*) - P_F(q)) \ge 0.$$

Since we are given $P_F(q_*) = \alpha$ and $P_F(q) \le \alpha$, this implies $P_D(q_*) \ge P_D(q)$, as desired.

It remains to show that q_* exists. Note that

$$\begin{aligned} P_F(q_*) &= \int q_*(y) p_{Y|H}(y|H_0)\mathrm{d}y \\ &= \mathbb{P}[L(y) > \eta | H = H_0] + p\mathbb{P}[L(y) = \eta | H = H_0]. \end{aligned}$$

We would like to force this quantity to be equal to α, so we can set

$$p = \frac{\alpha - \mathbb{P}[L(y) > \eta | H = H_0]}{\mathbb{P}[L(y) = \eta | H = H_0]},$$

which works as long as we choose η correctly. In particular, since we want $p \in [0, 1]$, we need

$$\alpha \leq \mathbb{P}[L(y) > \eta | H = H_0] + \mathbb{P}[L(y) = \eta | H = H_0].$$

One possibility is to choose η as the smallest possible number satisfying $\mathbb{P}[L(y) > \eta | H = H_0] \leq \alpha$, so that the above inequality holds as soon as we add in the point mass at $L(y) = \eta$.

This construction works, so we are mostly done. The only edge case is when $\mathbb{P}[L(y) = \eta | H = H_0] = 0$, since this would force

$$\frac{\alpha - \mathbb{P}[L(y) > \eta | H = H_0]}{\mathbb{P}[L(y) = \eta | H = H_0]} \to \infty.$$

In this case we have

$$\alpha \leq \mathbb{P}[L(y) > \eta | H = H_0] + 0 \leq \alpha,$$

so $\mathbb{P}[L(y) > \eta] = \alpha$ and $P_F(q_*) = \mathbb{P}[L(y) > \eta | H = H_0] = \alpha$ irrespective of the chosen value of p. □

4.3 Efficient Frontier of Operating Points

The Neyman-Pearson function satisfies a few nice properties.

Corollary 4.2

The Neyman-Pearson function $\zeta_{NP}(\cdot)$ is concave.

Proof. This is a direct result of the example above. For any two points on the efficient frontier, we can achieve at least the line connecting these two points by

randomizing between the two hypotheses corresponding to these points. Thus, the function is concave. □

Corollary 4.3

Let η_0 be the smallest number satisfying

$$\mathbb{P}[L(y) < \eta_0 | H = H_0] \leq \alpha.$$

Then

$$\dot{\zeta}_{NP}(P_F(\eta_0)) = \eta_0.$$

Proof. Define $p_{L|H}(\ell|H_0) = \mathbb{P}[L(y) = \ell | y \sim p_{Y|H}(y|H_0)]$. Now,

$$\begin{aligned} P_D(\eta) &= \int \mathbb{1}(L(y) \geq \eta) L(y) p_{Y|H}(y|H_0) \mathrm{d}y \\ &= \mathbb{E}_{y \sim p_{Y|H}(y|H_0)}[\mathbb{1}(L(y) \geq \eta) L(y) | H = H_0] \\ &= \mathbb{E}_{\ell \sim p_{L|H}(\ell|H_0)}[\mathbb{1}(\ell \geq \eta)\ell | H = H_0] \\ &= \int_\eta^\infty \ell p_{L|H}(\ell|H_0) \mathrm{d}\ell. \end{aligned}$$

Taking a derivative wrt ℓ gives

$$\frac{\mathrm{d}P_D(\eta_0)}{\mathrm{d}\ell} = -\eta_0 p_{L|H}(\ell|H_0).$$

Applying the same argument to P_F gives

$$\frac{\mathrm{d}P_F(\eta_0)}{\mathrm{d}\ell} = -p_{L|H}(\ell|H_0).$$

We showed that η_0 was the "good" hypothesis in the proof of the Neyman-Pearson lemma, so $\zeta_{NP}(P_F(\eta_0)) = P_D(\eta_0)$. Thus,

$$\dot{\zeta}_{NP}(P_F(\eta_0)) = \frac{\dot{P}_D(\eta_0)}{\dot{P}_F(\eta_0)} = \eta_0,$$

as desired. □

5 March 5, 2024

5.1 Jensen's Inequality

Theorem 5.1

If $\phi(\cdot)$ is concave and $v \in \mathcal{V}$ for some alphabet $\mathcal{V}$, then

$$\mathbb{E}[\phi(v)] \leq \phi(\mathbb{E}[v]).$$

Here we say "concave" in the sense of a parabola that opens **downwards**. Intuitively, the expected value of a set of points on such a parabola (i.e., their y-values) is less than the expected value of the point on the parabola that has the expected x-values of the points, because the edge points on the parabola drags down the expectation.

Proof. We can prove with induction on the size of $\mathcal{V}$. When $|\mathcal{V}| = 2$, we have

$$p_v(v_1)\phi(v_1) + p_v(v_2)\phi(v_2) \leq \phi(p_v(v_1)v_1 + p_v(v_2)v_2),$$

directly from concavity. Now, for any larger $|\mathcal{V}|$,

$$\begin{aligned}
\sum_i p_v(v_i)\phi(v_i) &= p_v(v_1)\phi(v_1) + \sum_{i>1} p_v(v_i)\phi(v_i) \\
&= p_v(v_1)\phi(v_1) + (1 - p_v(v_1)) \sum_{i>1} \frac{p_v(v_i)}{1 - p_v(v_1)} \phi(v_i) \\
&\leq p_v(v_1)\phi(v_1) + (1 - p_v(v_1))\phi\left(\sum_{i>1} \frac{p_v(v_i)v_i}{1 - p_v(v_1)} \right) \\
&\leq \phi(\sum p_v(v_i)v_i) = \phi(\mathbb{E}[v]),
\end{aligned}$$

assuming without loss of generality that $\phi(v_1) \neq 1$. If there doesn't exist such a v_i, then all of the $p_v(v_i) \in \{0, 1\}$, so $\mathbb{E}[\phi(v)] = \phi(v_j) = \phi(\mathbb{E}[v])$ for the singular v_j with $p_j = 1$. □

5.2 Csiszar's Inequality

Theorem 5.2

Given positive finite-length sequences $a_1,\dots,a_N$ and $b_1,\dots,b_N$, and strictly convex f,

$$\sum_{i=1}^{N} b_i f\left(\frac{a_i}{b_i}\right) \geq \left(\sum_{i=1}^{N} b_i\right) f\left(\frac{\sum_{i=1}^{N} a_i}{\sum_{i=1}^{N} b_i}\right),$$

with equality if and only if a_i/b_i is constant.

Proof. This is essentially a direct application of Jensen's inequality. To convert the LHS into a probability distribution, we divide out by $\sum b_i$, and then apply Jensen's to finish:

$$\begin{aligned}\sum_{i=1}^{N} b_i f\left(\frac{a_i}{b_i}\right) &= \sum b_i \sum_{i=1}^{N} \frac{b_i}{\sum b_i} f\left(\frac{a_i}{b_i}\right) \\ &\geq \sum b_i f\left(\sum_{i=1}^{n} \frac{b_i}{\sum b_i} \cdot \frac{a_i}{b_i}\right) \\ &= \sum b_i f\left(\frac{\sum a_i}{\sum b_i}\right).\end{aligned}$$

□

5.3 Gibbs' Inequality

This is a very important inequality.

Theorem 5.3

Let v be an r.v. distributed as p. For any alternative distribution q,

$$\mathbb{E}_p[\log p(v)] \geq \mathbb{E}_p[\log q(v)],$$

with equality iff $q \equiv p$.

Proof. By Jensen,

$$\begin{aligned}\mathbb{E}_p[\log q(v)] - \mathbb{E}_p[\log p(v)] &= \mathbb{E}_p\left(\log \frac{q(v)}{p(v)}\right) \\ &\leq \log \mathbb{E}_p\left(\frac{q(v)}{p(v)}\right) = 0.\end{aligned}$$

□

6 March 7, 2024

6.1 Complete Data

Consider the following setup. We have some observed data y sampled from Y, which has distribution $p_Y(\cdot; x)$ for some $x \in \mathcal{X}$. We want

$$\hat{x}(y) = \arg\max_{a \in \mathcal{X}} \log p_Y(y; a),$$

i.e., the maximum likelihood hypothesis that explains the data we observed. Define $\ell_Y(a; y) = \log p_Y(y; a)$, so that

$$\hat{x}(y) = \arg\max_{a \in \mathcal{X}} \ell_Y(a; y).$$

Now suppose that there exists some latents z sampled from Z, which has distribution $p_Z(\cdot; x)$. Z is the "complete data", i.e., $Y = g(Z)$ for some deterministic g. Z always exists, but we may not be able to observe it; we have to guess a form that is reasonable and hope that the results work. Given the data, we can compute an expected likelihood of our latent distribution:

$$\mathbb{E}_{Z|Y}(\cdot|y; x')[\ell_Z(x; z)],$$

for any $x' \in \mathcal{X}$. We cannot compute the true value, since we are assuming that we don't have access to the complete data (in practice, the "complete data" is just a guess, e.g., I assume that the data is being generated from k-clusters). Also, since $Y = g(Z)$ deterministically, we can say that

$$p_Z(z; x) = p_{Z,Y}(z, y; x) = p_{Z|Y}(z|y; x)p_Y(y; x),$$

so taking the log and then expectation of both sides gives

$$\log p_Y(y;x) = \mathbb{E}_{p_{Z|Y}(\cdot|y;x')}[\log p_Z(z;x)] - \mathbb{E}_{p_{Z|Y}(\cdot|y;x')}[\log p_{Z|Y}(z|y;x)]$$

for all $x, x' \in \mathcal{X}$. Re-write this term wise as

$$\ell_Y(x;y) = U(x;x') + V(x;x').$$

By Gibbs, $V(x;x') \geq V(x';x')$, so this rearranges to

$$\ell_Y(x;y) \geq (U(x;x') - U(x';x')) + U(x';x') + V(x';x') \geq \Delta(x,x') + \ell_Y(x';y).$$

Given that we choose x s.t. $\Delta(x,x') > 0$, we now have $\ell_Y(x;y) > \ell_Y(x';y)$, which gives a way to guarantee an increase in log likelihood. This is the foundation for the EM-algorithm.

6.2 Expectation-Maximization Algorithm

The EM-algorithm is as follows:

- Initialize $t = 0$ and a guess for $\hat{x}^{(0)}$.
- **E**: Compute
$$U(x;\hat{x}^{(t)}) = \mathbb{E}_{p_{Z|Y}(\cdot|y;\hat{x}^{(t)})}[\log p_Z(z;x)].$$
- **M**: Compute
$$\hat{x}^{(t+1)} = \arg\max_{x \in \mathcal{X}} U(x;\hat{x}^{(t)}).$$
- Increment t and repeat the **EM** cycle until convergence.

The intuition behind the **M**-step here is to guarantee that $\Delta(\hat{x}^{(t+1)}, \hat{x}^{(t)}) \geq 0$, so that we have an increasing (non-decreasing) sequence of likelihoods

$$\{\ell_Y(\hat{x}^{(t)})\}_{t \geq 0}.$$

The hope is that after enough steps, we can converge on a optimal value.

6.3 EM on Logistic Regression

The goal of linear regression is to assign a probability $\sigma(\theta^T x)$ that the assigned label of a data point is 1. Suppose we have complete data

$$\mathcal{D}_+ = \{(u_i, v_i, w_i)\}_{1 \le i \le N},$$

where the hidden part of the data is a mixture component $w_i \in \{0, 1\}$. The observed data is

$$\mathcal{D} = \{(u_i, v_i)\}_{1 \le i \le N},$$

which are the standard pairs of vectors u_i and labels $v_i \in \{\pm 1\}$ for logistic regression. For the E step, we are interested in $p_Z(z)$, which is governed by

$$p_{V,U,W}(v, u, w) = p_{V|U,W}(v|u, w)p_U(u)p_W(w).$$

We have

$$p_{V|U,W}(v|u, w) = \frac{1}{1 + e^{-v x_w^T t(u)}},$$

where $t(u)$ is a feature vector extracted from u, and x_w is the learned hypothesis vector given mixture w. If we suppose that $w \sim \textsc{Bern}(q)$ for unknown q, then we also have

$$p_W(w) = \exp\left(w \ln\left(\frac{q}{1-q}\right) + \ln(1-q)\right),$$

so

$$\ln p_Z(z) = -\ln(1 + e^{-v x_w^T t(u)}) + \ln p_U(u) + w \ln\left(\frac{q}{1-q}\right) + \ln(1-q).$$

The full hypothesis that we want to learn is $\theta = (x_0, x_1, q)$. Now we can compute the expectation. For the first term,

$$\begin{aligned}
&\mathbb{E}_{p_{Z|Y}(\cdot|y;\theta')}[-\ln(1 + e^{-v x_w^T t(u)})] \\
&\quad = -p_{W|U,V}(0|u, v; \theta')\ln(1 + e^{-v x_0^T t(u)}) - p_{W|U,V}(1|u, v; \theta')\ln(1 + e^{-v x_1^T t(u)}).
\end{aligned}$$

For the last term,

$$\mathbb{E}_{p_{Z|Y}(\cdot|y;\theta')}[w \ln\left(\frac{q}{1-q}\right) + \ln(1-q)] = p_{W|U,V}(1|u, v; \theta')\ln\left(\frac{q}{1-q}\right) + \ln(1-q),$$

so the whole expectation is

$$\begin{aligned}
U(\theta,\theta') &= \mathbb{E}_{p_{Z|Y}(\cdot|y;\theta')}[\sum_{i=1}^{N} \ln p_Z(z_i)] \\
&= -\sum_{i=1}^{N} p_{W|U,V}(0|u_i,v_i;\theta') \ln(1+e^{-v_i x_0^T t(u_n)}) \\
&\quad -\sum_{i=1}^{N} p_{W|U,V}(1|u_i,v_i;\theta') \ln(1+e^{-v_i x_1^T t(u_n)}) \\
&\quad +\sum_{i=1}^{N} \ln p_U(u_i) + \sum_{i=1}^{N} p_{W|U,V}(1|u_i,v_i;\theta') \ln\left(\frac{q}{1-q}\right) + N\ln(1-q).
\end{aligned}$$

To calculate each of the posteriors, we can use Bayes':

$$\begin{aligned}
&p_{W|U,V}(1|u_i,v_i;\theta') \\
&\quad = \frac{p_W(1;\theta')p_{V|U,W}(v_i|u_i,w_i=1;\theta')}{p_W(0;\theta')p_{V|U,W}(v_i|u_i,w_i=1;\theta') + p_W(1;\theta')p_{V|U,W}(v_i|u_i,w_i=1;\theta')} \\
&\quad = \frac{q'(1+\exp(-v_i(x_1')^T t(u_i)))^{-1}}{(1-q')(1+\exp(-v_i(x_0')^T t(u_i)))^{-1} + q'(1+\exp(-v_i(x_1')^T t(u_i)))^{-1}} \\
&\quad = \left(1+\frac{1-q'}{q'}\frac{1+\exp(-v_i(x_1')^T t(u))}{1+\exp(-v(x_0')^T t(u))}\right)^{-1}.
\end{aligned}$$

For the **M** step, we can optimize each component in $\theta = (x_0, x_1, q)$ separately. In particular, we can take

$$x_i^{(t+1)} = \arg\min_x \sum_{i=1}^{N} \ln(1+e^{-v_i x^T t(u_i)})$$

and

$$q^{(t+1)} = \arg\max_q \sum_{i=1}^{N} p_{W|U,V}(1|u_i,v_i;\theta') \ln\left(\frac{q}{1-q}\right) + N\ln(1-q).$$

The posteriors rely only on q', so they are constant relative to our maximization. We can solve directly for the maximization for $q^{(l+1)}$:

$$\left(\frac{1}{N}\sum_{i=1}^{N} p_{W|U,V}(1|u_i,v_i;\theta')\right)\ln\left(\frac{q}{1-q}\right)+\ln(1-q)$$
$$=\left(\frac{1}{N}\sum_{i=1}^{N} p_{W|U,V}(1|u_i,v_i;\theta')\right)\ln(q)+\left(1-\frac{1}{N}\sum_{i=1}^{N} p_{W|U,V}(1|u_i,v_i;\theta')\right)\ln(1-q).$$

This the expected value of log Bernoulli $\ln q$ with respect to another Bernoulli, so by Gibbs,

$$q^{(t+1)}=\frac{1}{N}\sum_{i=1}^{N} p_{W|U,V}(1|u_i,v_i;\theta').$$

6.4 EM on Infectious Disease

Say we have a population of N people over T years, and data $y_i \in \{0,1\}$ signalling whether a subject $i \in [N]$ contracted an infectious disease by te end of the T years. We have access to a record $\mu_{ij} \in \{0,1\}$ that tells us whether or not subject $i \in [N]$ was exposed to the disease in year $j \in [T]$. Our goal is to estimate the infection rate r_j for each year $j \in [T]$.

Naturally, the "complete" data that explains our observations can be represented by $z_{ij} \in \{0,1\}$ indicating whether i contracted the disease during year j. We don't have access to this data, but we can use the EM algorithm on this latent space to estimate r_j.

We can show using a bit of casework that

$$p_{z_{ij}}(z_{ij};r)=r_j^{\mu_{ij}z_{ij}}(1-r_j)^{\mu_{ij}z_{ij}},$$

so

$$\ln p_Z(z;r)=\sum_{i=1}^{N}\sum_{j=1}^{T} z_{ij}\mu_{ij}\ln r_j+\mu_{ij}(1-z_{ij})\ln(1-r_j).$$

Now we perform **E**:

$$\mathbb{E}_{p_{Z|Y}(\cdot|y;r')}\left[\sum_{i=1}^{N}\sum_{j=1}^{T} z_{ij}\mu_{ij}\ln r_j + \mu_{ij}(1-z_{ij})\ln(1-r_j)\right]$$
$$= \sum_{j=1}^{T}\sum_{i=1}^{N} \mu_{ij}\mathbb{E}_{p_{Z|Y}(\cdot|y;r')}[z_{ij}]\left(\frac{r_j}{1-r_j}\right) + \mu_{ij}\ln(1-r_j).$$

We also need the expectation for each z_{ij}:

$$\mathbb{E}_{p_{Z|Y}(\cdot|y;r')}[z_{ij}] = \mathbb{P}[z_{ij}=1|Y=y,r'] = \frac{\mathbb{1}(y_i=1)\mu_{ij}r'_j}{\sum_{j'=1}^{T}\mathbb{1}(y_i=1)\mu_{ij'}r'_{j'}}.$$

To perform **M**, note that we can optimize one component at a time by taking the inner sum, i.e.,

$$r_j = \arg\max_{r_j}\left(\sum_{i=1}^{N}\mu_{ij}\mathbb{E}_{p_{Z|Y}(\cdot|y;r')}[z_{ij}]\left(\ln\frac{r_j}{1-r_j}\right) + \mu_{ij}\ln(1-r_j)\right).$$

Taking critical points gives

$$\sum_{i=1}^{N}\left(\mu_{ij}\mathbb{E}_{p_{Z|Y}(\cdot|y;r')}[z_{ij}]\cdot\frac{1-r_j}{r_j}\frac{(1-r_j)+r_j}{(1-r_j)^2} - \frac{\mu_{ij}}{1-r_j}\right) = 0,$$

which gives

$$r_j = \frac{\sum_{i=1}^{N}\left(\mu_{ij}\cdot\frac{\mathbb{1}(y_i=1)\mu_{ij}r'_j}{\sum_{j'=1}^{T}\mathbb{1}(y_i=1)\mu_{ij'}r'_{j'}}\right)}{\sum_{i=1}^{N}\mu_{ij}}.$$

Note that this result intuitively makes sense; we want r_j to approach the true proportion $\mathbb{E}[z_{ij}]$ of infected subjects, and we weight only by $\mu_{ij}=1$ because we can say definitively that $z_{ij}=1$ should only happen when subjects are exposed.

6.5 EM Alternate Formulation

Here we introduce an alternate formulation for the EM algorithm that we introduced earlier. Say the complete data is $Z \in \mathcal{Y}\times\mathcal{S}$, where $Y \in \mathcal{Y}$ is our observed data

and $S \in \mathcal{S}$ is the latent. Consider the objective function

$$\ln \tilde{p}_Y(y; x, q_{S|Y}(\cdot|y)) = \mathbb{E}_{q_{S|Y}(\cdot|y)}[\ln p_{Y,S}(y, s; x)] - \mathbb{E}_{q_{S|Y}(\cdot|y)}[\ln q_{S|Y}(s|y)],$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. This is similar to the original setup, but now the conditioning on x no longer affects the distribution over the expected values, and moreover the distribution over the expected values is an additional input to our likelihood function. I'm not sure what the significance of these differences are yet, but this setup will supposedly be important when we learn about variational inference.

Since we now have two inputs conditioning our likelihood function, we can alternate between them for our algorithm. First, guess $\hat{x}^{(0)}$, and then alternate

$$\hat{q}_{S|Y}^{(t)}(\cdot|y) = \arg\max_{q \in \mathscr{P}_+^{\mathcal{S}}} \ln \tilde{p}_Y(y; \hat{x}^{(t)}, q_{S|Y}(\cdot|y)),$$

$$\hat{x}^{(t+1)} = \arg\max_{x \in \mathcal{X}} \ln \tilde{p}_Y(y; x, \hat{q}_{S|Y}^{(t)}(\cdot|y)).$$

We will show that this is equivalent to the original EM formulation. First, note that

$$\begin{aligned}
\ln \tilde{p}_Y(y; x, q_{S|Y}(\cdot|y)) &= \mathbb{E}_{q_{S|Y}(\cdot|y)}[\ln p_{Y,S}(y, s; x)] - \mathbb{E}_{q_{S|Y}(\cdot|y)}[\ln q_{S|Y}(s|y)] \\
&= \mathbb{E}_{q_{S|Y}(\cdot|y)}[\ln p_Y(y; x)] + \mathbb{E}_{q_{S|Y}(\cdot|y)}[\ln p_{S|Y}(\cdot|y; x)] - \mathbb{E}_{q_{S|Y}(\cdot|y)}[\ln q_{S|Y}(s|y)] \\
&\leq \ln p_Y(y; x) + \mathbb{E}_{q_{S|Y}(\cdot|y)} \ln q_{S|Y}(\cdot|y; x) - \mathbb{E}_{q_{S|Y}} \ln q_{S|Y}(\cdot|y; x) = \ln p_Y(y; x),
\end{aligned}$$

by Gibbs. We have equality when $q_{S|Y}(\cdot|y) = p_{S|Y}(\cdot|y; x)$, so

$$q_{S|Y}^{(t)}(\cdot|y) = p_{S|Y}(s|y; x^{(t)}).$$

Now,

$$\ln(\tilde{p}_y; x, q_{S|Y}^{(t)}(\cdot|y)) = \mathbb{E}_{p_{S|Y}(\cdot|y;x^{(t)})}[\ln p_{Y,S}(y, s; x)] - \mathbb{E}_{p_{S|Y}(\cdot|y;x^{(t)})}[\ln p_{S|Y}(s|y)].$$

The second term on the RHS is not a function of x, so it doesn't affect our arg max. Since $Z = (Y, S)$, we thus have

$$\hat{x}^{(t+1)} = \arg\max_{x \in \mathcal{X}} \mathbb{E}_{p_{Z|Y}(\cdot|y;x^{(t)})}[\ln p_Z(z; x)] = U(x; \hat{x}^{(t)}),$$

which recovers the original EM form, so we are done.

7 March 12, 2024

7.1 Generalized Cost Functions

Today we consider more generalized Bayes Decision models that outputs a decision $q : \mathcal{X} \to \mathbb{R}$ which is a probability distribution over the hypothesis space, rather than a single output.

To generalize metrics for the effectiveness of such models, we need to generalize the cost functions that we've been using. Our new cost functions must be functions of the correct hypothesis and the distribution output of the model, i.e., $C : (\mathcal{X}, (\mathcal{X} \to \mathbb{R})) \to \mathbb{R}$. There are certain qualities that we want these cost functions to have.

Definition 7.1

A cost function $C(\cdot, \cdot)$ is **proper** if it leads to us choosing the true belief, i.e.,

$$p_{X|Y}(\cdot|y) = \arg\min_q \mathbb{E}[C(x, q)|Y = y] \quad \forall y \in \mathcal{Y}.$$

Claim 7.2

The log-loss cost criterion is proper.

Proof.

$$\mathbb{E}_{p_{X|Y}(\cdot|y)}[-A \log q(x) + B(x)]$$

is minimized when $q = p_{X|Y}$, by Gibbs. □

Definition 7.3

A cost function $C(\cdot, \cdot)$ is **local** if there exists $\phi : (\mathcal{X}, \mathbb{R}) \to \mathbb{R}$ s.t. $C(x, q) = \phi(x, q(x))$ for all $x \in \mathcal{X}$.

In words, local cost functions are only functions of the probability of the actual outcome x, rather than the entire probability distribution.

Theorem 7.4

Given alphabet $\mathcal{X}$ with size $L := |\mathcal{X}| \geq 3$, log-loss is the only smooth, local, proper cost function.

Proof. Let $\mathcal{X} = \{x_1, \ldots, x_L\}$, $q_l = q(x_l)$,$p_l = p_{X|Y}(x_l|y)$, and $\phi_l(\cdot) = \phi_l(x_l, \cdot)$ be our local cost functions. Since the cost functions are proper,

$$(p_1, \ldots, p_L) = \underset{q_i \text{ valid}}{\arg\min} \sum_{l=1}^{L} p_l \phi_l(q_l).$$

We can use Lagrange multipliers to □

8 March 19, 2024

8.1 Continuous Information Theory

Many of the formulas we used in the discrete case work mostly the same in the continuous case. Let X, Y be continuous r.v.s; then,

$$h(X) = -\int_{\infty}^{\infty} p_X(x) \log p_X(x) \mathrm{d}x,$$

and

$$h(X|Y = y) = -\int_{\infty}^{\infty} p_{X|Y}(x|y) \log p_{X|Y}(x|y) \mathrm{d}x,$$

and

$$h(X|Y) = \int_{-\infty}^{\infty} p_Y(y) h(X|Y = y) \mathrm{d}y = -\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{X,Y}(x, y) \log p_{X|Y}(x|y) \mathrm{d}x \mathrm{d}y.$$

Mutual information is defined the same

$$I(X; Y) = h(X) - h(X|Y) = h(Y) - h(Y|X),$$

and information divergence as well

$$D(p||q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x.$$

There are some differences in the way that these functions behave under coordinate transformations. In the discrete cases, relabelling the alphabet does not change the behavior of any functions. In the continuous case, suppose that $X = g(S)$ for some monotonically increasing, differentiable mapping on S. Then for any x and

$s = g^{-1}(x)$, we must have:

- $\mathrm{d}x = \mathrm{d}s\dot{g}(s)$
- $p_X(x)\dot{g}(s) = p_S(s)$. This isn't related to the main discussion, but there are a few ways to think about why this must be the case. The first is that we want the unit integral area to be the same should remain the same under transformation, i.e., $p_X(x)\mathrm{d}x = p_S(s)\mathrm{d}s$. Another way to think about this is that an interval of unit length $[s, s+\mathrm{d}s]$ gets mapped to $[g(s), g(s)+\dot{g}(s)\mathrm{d}s]$, which is longer by a factor of $\dot{g}(s)$; therefore, the height $p_X(x)$ should be shorter by a factor of $\dot{g}(s)$, in order to keep the total area 1.

So the differential entropy is

$$h(X) = -\int_{\infty}^{\infty} \frac{p_S(s)}{\dot{g}(s)} \log \frac{p_S(s)}{\dot{g}(s)} \dot{g}(s)\mathrm{d}s = h(S) + \mathbb{E}[\log(\dot{g}(s))].$$

It is not invariant to coordinate transformations. On the other hand, mutual information under coordinate transformation is

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= h(S) + \mathbb{E}[\log(\cdot g(s))] - h(S|Y) - \mathbb{E}[\log(\cdot g(s))] \\ &= h(S) - h(S|Y), \end{aligned}$$

so it remains the same. Similarly, information divergence

$$\begin{aligned} D(p_x \| q_x) &= \int_{-\infty}^{\infty} p_X(x) \log \frac{p_X(x)}{q_X(x)} \mathrm{d}x \\ &= \int_{-\infty}^{\infty} p_S(s) \log \frac{p_S(s)}{q_S(s)} \mathrm{d}s = D(p_s \| q_s) \end{aligned}$$

is also invariant.

8.2 Gaussian Distribution Information Measures

Now we derive some information measures on Gaussian distributions. Let $Y = aX + Z$, where $X \sim \mathcal{N}(0,1)$ and $Z \sim \mathcal{N}(0,1)$. This is the scalar version of the so-

called **innovations form** for Gaussian vectors. Then, we have that

$$\begin{aligned} p_{X|Y}(x|y) &\propto \exp\left(-\frac{1}{2}(y-ax)^2\right)\exp\left(-\frac{1}{2}x^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left(x^2(a^2+1)-2ayx+y^2\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{x-ay/(a^2+1)}{1/(a+1)}\right)^2\right), \end{aligned}$$

so

$$p_{X|Y}(x|y) \sim \mathcal{N}(\mu_{X|Y}, \lambda_{X|Y}) = \mathcal{N}\left(\frac{ay}{a^2+1}, \frac{1}{a+1}\right).$$

Example 8.1

Differential entropy.

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then,

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} p_X(x)\log p_X(x)\mathrm{d}x \\ &= -\int_{-\infty}^{\infty} p_X(x)\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2} \\ &= \frac{1}{2}\log(2\pi e\sigma^2). \end{aligned}$$

Example 8.2

Conditional differential entropy.

Consider the example (X, Y) defined from before. From the conditional distribution,

$$h(X|Y=y) = \frac{1}{2}\log\left(\frac{2\pi e}{a+1}\right).$$

Therefore,

$$h(X|Y) = \int_{-\infty}^{\infty} h(X|Y=y)\mathcal{N}(y; 0, a^2+1)\mathrm{d}y = \frac{1}{2}\log\left(\frac{2\pi e}{a+1}\right).$$

Example 8.3

Mutual information.

Using the same example (X, Y) defined above,

$$I(X;Y) = h(X) - h(X|Y) = \frac{1}{2}\log(a+1).$$