Machine Learning Notes

Matthew Scicluna

December 25, 2017

Contents

1 Statistics, Probability and Optimization Background **3**
- 1.1 Useful Statistical Ideas . . . 3
  - 1.1.1 Empirical Distribution . . . 3
- 1.2 Bayesian Statistics . . . 5
  - 1.2.1 Existence Of The Prior . . . 5
  - 1.2.2 Likelihood Principle . . . 5
- 1.3 Optimization . . . 7
  - 1.3.1 Langrangian Duality . . . 7
  - 1.3.2 KKT Conditions . . . 8

2 Statistical Decision Theory **10**
- 2.1 Formal Set-Up . . . 10
- 2.2 Procedure Analysis . . . 10
  - 2.2.1 Frequentist Risk Perspective . . . 10
  - 2.2.2 Bayesian Risk Perspective . . . 11
- 2.3 Types of Procedures . . . 12
  - 2.3.1 Parameter Estimation . . . 12
  - 2.3.2 Prediction . . . 13
  - 2.3.3 Regression . . . 13

3 Information Theory **15**
- 3.1 Entropy . . . 15
- 3.2 KL Divergence . . . 16
- 3.3 Mutual Information . . . 17
- 3.4 Differential Entropy . . . 18
- 3.5 Entropy and Estimation . . . 18
  - 3.5.1 Maximum Likelihood Estimation . . . 19
  - 3.5.2 Maximum Entropy Principle . . . 19
  - 3.5.3 MaxENT and the Exponential Family . . . 22

1 Statistics, Probability and Optimization Background

We discuss the statistical background of Machine Learning and introduce the ideas that will be used throughout these notes.

1.1 Useful Statistical Ideas

1.1.1 Empirical Distribution

Given some data $x_1, \cdots x_n \sim F$ where F is an unknown CDF, we want to approximate this using some mapping $\hat{F}$ called the **Empirical Distribution** of the data.

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq t\}} \tag{1}$$

It can be shown that $\hat{F}(t) \to F(t)$ *a.s.* $\forall t$, justifying its use as an approximation of F, provided enough data has been observed. As with F we can approximate f. We define the **Empirical Density Function** $\hat{f}$:

$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i, t) \tag{2}$$

Where δ is defined differently in the continuous and discrete case. In the continuous case it is called the **Dirac Delta Function**:

$$\delta(x, y) = \begin{cases} \infty & x = y \\ 0 & \text{o.w.} \end{cases} \tag{3}$$

Additionally, we suppose that:

1. $\int_{-\infty}^{\infty} \delta(t, y) dt = 1$
2. $\int \delta(t, y) f(t) dt = f(y)$, for any f with compact support that is continuous around y

This is not a function, but is called a *Generalized Function*. In the discrete case things are much simpler, as we can use the simpler **Kronecker delta function**:

$$\delta(x, y) = \begin{cases} 1 & x = y \\ 0 & \text{o.w.} \end{cases} \tag{4}$$

Finally, we notice that $\hat{f}$ and $\hat{F}$ satisfy an important relationship that would be expected from the cdf and pdf: $\int_{-\infty}^{t} \hat{f}(y) dy = \hat{F}(t)$.

$$
\begin{aligned}
\int_{-\infty}^{t} \hat{f}(y)dy &= \int_{-\infty}^{t} \frac{1}{n} \sum_{i=1}^{n} \delta(x_i, y)dy \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{t} \delta(x_i, y)dy \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{1}_{\{x_i \leq y\}} \delta(x_i, y)dy \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq y\}} \\
&= \hat{F}(t)
\end{aligned}
$$

1.2 Bayesian Statistics

Given an Event, what does the probability of that Event *mean*? There are two interpretations:

1. **Frequentists**: the *limiting frequency* of the event
2. **Bayesians**: the *reasonable expectation* that the event occurs

The *reasonable expectation* can be further broken down into two views. The **Objective Bayesians** view the *reasonable expectation* as the *state of knowledge*. They view probability as an extension of propositional logic, which is described in [1]. The **Subjective Bayesians** view probability as a quantification of *personal belief*. The main difference between the groups is in how they choose their priors: the Subjective Bayesians use knowledge about or prior experience with model parameters, wheras the Objectivists try to introduce as little prior knowledge as possible, using noninformative priors.

1.2.1 Existence Of The Prior

We need a theoretical justification for why we assume the existence of a prior distribution on θ in the first place! The justification for this requires the **Infinite Exchangeable** assumption of the data $\{x_i\}_{i=1}^{\infty}$. This is satisfied when, given a sequence of random variables, any finite subset $\{x_j\}_{j=1}^{n}$, and any permutation of this subset $\pi_{1:n}$

$$p(x_1, \cdots, x_n) = p(x_{\pi_1}, \cdots, x_{\pi_n}) \tag{5}$$

It turns out the above is equivalent to assuming the existence of the prior! The following theorem makes this precise.

Theorem 1.1 (De Finetti Theorem). *A sequence is Infinite Exchangeable iff for any n*

$$p(x_1, \cdots x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) P(d\theta)$$

for some measure P on θ. Also, if θ has a density then $P(d\theta) = p(\theta)d\theta$. Note: θ may be infinite!

This theory says that, if we assume exchangable data (and iid $\Rightarrow$ exchangeable), then there must exist a θ, $p(x|\theta)$ and distribution P on θ! So the idea of having a prior distribution on the parameters does have theory to back it up!

1.2.2 Likelihood Principle

The **Likelihood Principle** says that all the evidence in a sample relevent to parameters θ is contained in the likelihood function. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other.[2] This principle, if you believe it, gives a good justification for Bayesianism since $p(\theta|D) \propto P(D|\theta)P(\theta)$ (i.e. θ is being inferred from the likelihood).

Frequentists do not like the Likelihood Principle as it leads to contradictions with their methodology - see the following coin tossing example [3].

If you have trouble accepting the Likelihood Principle, it has been shown to be equivalent to two milder principles:

1. **Sufficiency Principle**: If two different observations x, y are such that $T(x) = T(y)$ for sufficient statistic T, then inference based on x and y should be the same.

2. **Conditionality Principle**: If an experiment concerning inference about θ is chosen from a collection of possible experiments independently, then any experiment not chosen is irrelevant to the inference.

Of these, Sufficiency is accepted by both Frequentists and Bayesians, while the Conditionality principle is debated.

1.3 Optimization

We now discuss how to solve optimization problems. This section is largly based off of [4]. We want to minimize our **Objective Function** $f_0 : \mathcal{D} \to \mathbb{R}$ w.r.t some our **Optimization Variable** $x \in \mathbb{R}^n$ subject to some **Inequality Constraints** $f_i : \mathbb{R}^n \to \mathbb{R}$ and some **Equality Constraints** $h_i : \mathbb{R}^n \to \mathbb{R}$. Formally:

$$\begin{aligned} &\text{Minimize } f_0(x) \\ &\text{subject to } f_i(x) \leq 0 \ i = 1, \cdots, m \\ &\qquad\qquad h_i(x) = 0 \ i = 1, \cdots, p \end{aligned}$$

We call a point x **Feasible** if $x \in \mathcal{D}$ and satisfies the constraints. We call x **Strictly Feasible** if $x \in \text{int}(\mathcal{D})$ and satisfies $f_1(x) < 0, \cdots, f_m(x) < 0$ and $h_1(x) = 0, \cdots, h_p(x) = 0$. The **Optimal Value** of this problem is $p^* = \inf\{f_0(x) : \text{x satisfies constraints}\}$. We let $p* = \infty$ if there are no feasible points, and $p^* = -\infty$ is the problem is unbounded from below. If x feasible and $f_0(x) = p^*$ then we call it **Optimal**.

1.3.1 Langrangian Duality

We can solve the above problem using the **Lagrangian** $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ with domain $\mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \tag{6}$$

This problem may be difficult to solve. The problem can be simplified by introducing the **Lagrange dual function** $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) \tag{7}$$

We call (6) the **Primal Problem** and (7) the **Dual Problem**. We note that g is concave (regardless of f_0) and can be $-\infty$. We call λ, ν **Dual Feasible** if $\lambda \geq 0$ and $(\lambda, \nu) \in \text{dom}(g)$. We denote the **Dual Optimum** (supremum) of g as d^*. We now relate the primal and dual problems:

Theorem 1.2 (Lower Bound Property). *Let $\lambda \geq 0$, then $g(\lambda, \nu) \leq p^*$*

Proof. Note that for any feasible $\tilde{x}$ and $\lambda \geq 0$:

$$f_0(\tilde{x}) \geq L(\tilde{x}, \lambda, \nu) \geq \inf_{x \in D} L(x, \lambda, \nu) = g(\lambda, \nu)$$

and since p^* is the infemum of all feasible $\tilde{x}$, it follows that $p^* \geq g(\lambda, \nu)$ □

Instead of minimizing f_0 to get p^* we can maximize the lower bound g. This is an easier problem since g is concave. It is clear that we always have **Weak Duality** $p^* \geq d^*$, although we want **Strong Duality**: $p^* = d^*$. A sufficient condition for strong duality is **Slater's Condition**

Theorem 1.3 (Slater's Condition). *Suppose we have a convex primal (i.e. f_0, f_i convex and h_i affine), if there exists a strictly feasible x we then have strong duality*

1.3.2 KKT Conditions

We now state some conditions often used to determine whether a solution x^* of f is optimal. These are the **KKT Conditions**.

Theorem 1.4 (KKT Conditions). *Let f_0, f_i, h_i be differentiable. If x^* is optimal and (λ^*, ν^*) are dual optimal and we have strong duality; then the following Conditions must be satisfied:*

- $\nabla_x f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla_x f_i(x^*) + \sum_{i=1}^p \nu_i^* \nabla_x h_i(x^*) = 0 \rightarrow$ *Stationarity*
- $\lambda_i^* f_i(x^*) = 0 \rightarrow$ *Complementary Slackness*
- x^* *is feasible* $\rightarrow$ *Primal Feasibility*
- $\lambda^* \geq 0 \rightarrow$ *Dual Feasibility*

Proof. We have to show Stationarity and Complementary Slackness. Notice that

$$f_0(x^*) = g(\lambda^*, \nu^*) = \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu_i^* h_i(x) \right) \tag{8}$$

$$\leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \nu_i^* h_i(x^*) \tag{9}$$

$$\overset{(a)}{\leq} f_0(x^*) \tag{10}$$

Where (a) comes from the condition that $h_i(x^*) = 0 \Rightarrow \sum_{i=1}^p \nu_i^* h_i(x^*) = 0$ and $f_i(x^*) \leq 0, \lambda_i \geq 0 \Rightarrow \sum_{i=1}^m \lambda_i^* f_i(x^*) \leq 0$. We get that $\inf_x L(x, \lambda^*, \nu^*) = f_0(x^*)$, i.e. x^* minimizes $L(x, \lambda^*, \nu^*)$ which implies Stationarity. Complementary slackness comes from $\sum_{i=1}^m \lambda_i^* f_i(x^*) = 0$ and $\lambda_i^* f_i(x^*) \leq 0 \Rightarrow \lambda_i^* f_i(x^*) = 0$ for every i. □

Theorem 1.5. *If x^*, λ^*, ν^* satisfy the KKT conditions and the primal is convex; then x^* is optimal, λ^*, ν^* are dual optimal and we have strong duality.*

Proof.

$$g(\lambda^*, \nu^*) = \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu_i^* h_i(x) \right) \tag{11}$$

$$\overset{(a)}{=} f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \nu_i^* h_i(x^*) \tag{12}$$

$$\overset{(b)}{=} f_0(x^*) \tag{13}$$

Since $\lambda^* \geq 0$, $L(x, \lambda^*, \nu^*)$ is convex in x. By the KKT conditions, x^* is primal feasible and $\nabla_x L(x, \lambda^*, \nu^*)$ evaluated at x^* is $0 \Rightarrow x^*$ minimizes $L(x, \lambda^*, \nu^*)$ (a). (b) follows from complementary slackness. Therefore, we have strong duality and so x^* is optimal and λ^*, ν^* are dual optimal. □

So the KKT conditions are necessary for optimality under strong duality and sufficient if the primal is convex. Slater's condition can make this into an if and only if statement.

Theorem 1.6. *If Slaters condition is satisfied, then x^* is optimal if and only if there exists λ^*, ν^* that satisfy KKT conditions.*

2 Statistical Decision Theory

2.1 Formal Set-Up

We need a general framework to make data-driven decisions under uncertainty. More formally, we observe some **Data** $D \in \mathcal{D}$ which comes from some **Data Generating Distribution** $D \sim P$. Let $P \in \mathcal{P}$[1], where $\mathcal{P}$ is the set of possible distributions. Let $\mathcal{A}$ be our set of possible actions. To determine how good an action is, we define the **Loss** (cost) of doing that action as $L : \mathcal{P} \times \mathcal{A} \mapsto \mathbb{R}$. The goal is to determine a **Decision Rule** $\delta : \mathcal{D} \mapsto \mathcal{A}$ which, given data, produces an action.

Typically we consider $\mathcal{P}$ as a Parametric Family of distributions, and we use Θ interchangibly with $\mathcal{P}$, using that $P := P_\theta$.

2.2 Procedure Analysis

We need a way to assign a value to any δ and a way to compare these values to find which one is "best". One such way of doing this is via the Frequentist Risk.

2.2.1 Frequentist Risk Perspective

The first approach seeks to minimize the **Frequentist Risk**, which is defined as:

$$R(P, \delta) = \mathbb{E}_{D \sim P}\{L(P, \delta(D))\} \tag{14}$$

If we want to compare decision rules δ_1, δ_2 using def 14, we have to take into account P, since R varies with both P and δ. Sometimes one decision rule δ_1 is better then another δ_2 regardless of P, in which case we say δ_1 **Dominates** δ_2. More formally:

$$R(P, \delta_1) \leq R(P, \delta_2) \forall P \in \mathcal{P} \; and$$

$$\exists P \in \mathcal{P}, \; R(P, \delta_1) < R(P, \delta_2)$$

This basically means that δ_1 is a better decision rule then δ_2. Sometimes, there may be a "best" δ, one which isn't dominated by any other δ_0. We say δ is **Admissible** if $\nexists \delta_0$ s.t. δ_0 dominates δ. Note: we should rule out inadmissible decision rules (except for simplicity or efficiency) but not necessarily accept Admissible ones!

Unfortunately, different P's usually produce different optimal δ's! we must take into account the unknown P when minimizing (14). One way to take this into account is to use the **Minimax Criteria**: the optimal δ minimizes the Frequentist Risk in worst case scenerio.

$$\delta_{minimax} = \min_{\delta} \max_{P \in \mathcal{P}} \; R(P, \delta) \tag{15}$$

[1] Often P will describe an IID process, e.g. $D = (X_1, ..., X_n)$ where $X_i \overset{iid}{\sim} P_0$. In this case, the loss is usually written w.r.t P_0 instead of P.

If $\mathcal{P}$ is a Parameteric Family, we can handle the dependence of R on P by averaging it out, adding weights π over Θ to put more weight on certain θ's. We can then minimize over δ. This is called the **Bayes Risk**, even though it is Frequentist concept since it averages over D via (14).

$$\delta_{bayes} = \arg\min_{\delta} \int_{\Theta} R(P_\theta, \delta)\pi(\theta)d\theta \tag{16}$$

Where δ_{bayes} is called the **Bayes Rule**. Note that the Bayes Rule may not exist, and when they do they may not be unique.

2.2.2 Bayesian Risk Perspective

Note that (14) does not consider that we only observed one D. We can define a Risk function that does. The **Posterior Risk** is

$$R_B(\delta|D) = \int_{\Theta} L(P_\theta, \delta)p(\theta|D)d\theta \tag{17}$$

Where $p(\theta|D)$ is the posterior for a given prior $\pi(\theta)$. We can choose our decision rule based on this new risk function. This is called the **Bayes Estimator** or **Bayes Action** (not to be confused with the **Bayes Rule** above).

$$\delta_{post} = \arg\min_{\delta} R_B(\delta|D) \tag{18}$$

Notice that in (17), we do not consider different unobserved values of D, since the Bayesian would say they are irrelevent courtesy of the Conditionality Principle. For them, only the observed D matters for inference. Additionally, θ is integrated out in (17), meaning that (18) gives the undisputed optimal δ!

Note that the Frequentist can still use (18) by interpreting it as (16) with π as the "true" prior for Θ. We would then get that:

$$\begin{aligned}
\int_{\Theta} R(P_\theta, \delta)\pi(\theta)d\theta &= \int_{\Theta}\int_{D} L(P_\theta, \delta)P(D|\theta)P(\theta)dDd\theta \\
&\overset{(a)}{=} \int_{D}\int_{\Theta} L(P_\theta, \delta)P(\theta|D)P(D)d\theta dD \\
&= \int_{D} R_B(\delta|D)P(D)dD
\end{aligned}$$

Where (a) is due to Fubini's theorem (provided the integral is finite). It turns out that a *Bayes rule* can be obtained by taking the *Bayes action* for each particular D! See [5] for more details.

2.3 Types of Procedures

2.3.1 Parameter Estimation

Given a Parametric Family $\{P_\theta\}_{\theta\in\Theta}$, typically we have data $D = (X^{(1)}, \cdots, X^{(n)})$ where each $X^{(i)} \overset{iid}{\sim} P_\theta$. We want to use this data to estimate the true parameters θ. Hence, $\mathcal{A} = \Theta$ and $\delta(D)$ is some an **Estimator** of θ. The estimator should minimize the loss (more specifically the risk). One popular loss function is the **Squared Loss**: $L(\theta, \delta(D)) = ||\theta - \delta(D)||^2$. Note that since the data are IID we use the marginal density over X instead of the joint over D in the loss function.

If we take the expectation of the loss function above (the frequentist risk), we can decompose it nicely into two pieces:

$$\begin{aligned} R(P,\delta) &= \mathbb{E}_{D\sim P}\{||\theta - \delta(D)||^2\} \\ &= \mathbb{E}_{D\sim P}\{(\theta - \mathbb{E}_{D\sim P}\{\delta(D)\} + \mathbb{E}_{D\sim P}\{\delta(D)\} - \delta(D))^2\} \\ &= \underbrace{(\theta - \mathbb{E}_{D\sim P}\{\delta(D)\})^2}_{Bias^2} + \underbrace{\mathbb{E}_{D\sim P}\{(\mathbb{E}_{D\sim P}\{\delta(D)\} - \delta(D))^2\}}_{Variance} \end{aligned}$$

The above says that when we average this loss over all possible datasets, we can compare how much of the loss is due to the Bias and how much to the Variance of the estimator δ. This idea works for other loss functions, but the decomposition is not nearly as clean. Finally, this is a Frequentist idea since it involves taking an expectation over the data generating distribution, an idea doesn't appeal to Bayesians since it is contrary to the conditionality principle.

We conclude by showing an interesting result: for parameter estimation, the bayes action for the squared loss $\delta_{post}(D) = \mathbb{E}\{\theta|D\}$. This is a simple optimization problem:

$$\begin{aligned} R_B(\delta|D) &= \int_\Theta ||\theta - \delta(D)||^2 p(\theta|D)d\theta \\ &= \delta(D)^2 - 2\delta(D)\int_\Theta \theta p(\theta|D)d\theta + \int_\Theta \theta^2 p(\theta|D)d\theta \end{aligned}$$

and taking the derivative and setting to 0 yields:

$$\begin{aligned} \frac{\partial R_B}{\partial \delta} &= 2\delta(D) - 2\int_\Theta \theta p(\theta|D)d\theta = 0 \\ \Rightarrow \delta(D) &= \int_\Theta \theta p(\theta|D)d\theta = \mathbb{E}\{\theta|D\} \end{aligned}$$

2.3.2 Prediction

Let $D = ((X^{(1)}, Y^{(1)}), \cdots, (X^{(n)}, Y^{(n)}))$ where $X^{(i)} \in \mathcal{X}$ and $Y^{(i)} \in \mathcal{Y}$.

We put a density on X and Y: $(X^{(i)}, Y^{(i)}) \overset{iid}{\sim} P_{XY}$. Our action space $\mathcal{A} = \mathcal{Y}^{\mathcal{X}}$, the set of functions $f : \mathcal{X} \mapsto \mathcal{Y}$. Hence $\delta(D)$ is a **Learning Algorithm** which learns a function i.e. $\delta(D) = \hat{f}$.

We can evaluate the performance of f using a *prediction loss* $l : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$, a measure of the distance between a given prediction and its associated ground truth. We define the **Generalization Error** from l as:

$$L(P, f) = \mathbb{E}_{(X,Y)\sim P_{XY}}\{l(Y, f(X))\} \tag{19}$$

This is often called the Risk in Machine Learning. Note that we do not know (19), but we can approximate it using the below formula. This is called **Empirical Risk Minimization**

$$L(P, f) = \frac{1}{n}\sum_{i=1}^{n} l\left(Y^{(i)}, f(X^{(i)})\right) \tag{20}$$

Notice that (20) is just (19) with the empirical density substituted in place of the true, unknown one.

Prediction problems are called different things depending on whether $\mathcal{Y}$ is discrete or continuous. If $\mathcal{Y}$ discrete then the problem is called **Classification**, and if $\mathcal{Y}$ is not discrete e.g. $\mathcal{Y} = \mathbb{R}$, then it is referred to as **Regression**.

We previously described Prediction problems and noted that if $\mathcal{Y}$ is discrete then we refer to prediction as **Classification** and otherwise we refer to it as **Regression**.

2.3.3 Regression

We look at Regression. Let $D = ((X^{(1)}, T^{(1)}), \cdots, (X^{(n)}, T^{(n)}))$ with $X^{(i)} \in \mathcal{X}$ and $T^{(i)} \in \mathcal{T}$. Let $(X^{(i)}, T^{(i)}) \overset{iid}{\sim} P_{XT}$ and $l(t, y(x)) = |t - y(x)|^2$ (the squared loss). We need to find a function $y : \mathcal{X} \to \mathcal{T}$ which minimizes the Generalization Error:

$$\begin{aligned} L(P, y) &= \mathbb{E}_{(X,T)\sim P_{XT}}\{|t - y(x)|^2\} \\ &= \int\int |t - y(x)|^2 p(x, t) dx dt \end{aligned}$$

In this case we can find the optimal y by using calculus of variations. That is to say, the problem is reduced to an optimization problem. We denote $G(y, y', x) = \int |t - y(x)|^2 p(x, t) dt$ and use the Euler Lagrange equations to get that our stationary point must occur at

$$\begin{aligned} &\frac{\partial G(y, y', x)}{\partial y} - \frac{d}{dx}\frac{\partial G(y, y', x)}{\partial y'} = 0 \\ &\Rightarrow \frac{\partial G(y, y', x)}{\partial y} = 0 \end{aligned}$$

Since $\frac{\partial G(y,y',x)}{\partial y'} = 0$ since y' is not in G. We then solve:

$$\begin{aligned}\frac{\partial L(P,y)}{\partial y(x)} &= \frac{\partial}{\partial y(x)} \int |t - y(x)|^2 p(x,t) dt \\ &= 2 \int (t - y(x)) p(x,t) dt = 0\end{aligned}$$

Solving for the above we have that

$$\begin{aligned}\int (t - y(x)) p(x,t) dt = \int t p(x,t) dt - \int y(x) p(x,t) dt = 0 \\ \Rightarrow \int t p(x,t) dt = y(x) p(x) \\ \Rightarrow y(x) = \int t \frac{p(x,t)}{p(x)} dt = \mathbb{E}_{t \sim p(t|x)}\{t|x\}\end{aligned}$$

And so our learning algorithm $\delta(D)$ simply returns $y(x) = \mathbb{E}_{t \sim p(t|x)}\{t|x\}$. Of course, we don't know $p(t|x)$, so we would have to approximate it using the Empirical Risk Minimization described before!

The squared loss functions is a member of loss functions called the **Minkowski Loss**. This family has the following form:

$$L_q(P,y) = \int \int |t - y(x)|^q p(x,t) dx dt \tag{21}$$

We solve for the optimal $y(x)$ and set this to 0:

$$\frac{\partial L_q(P,y)}{\partial y(x)} = \int q|t - y(x)|^{q-1} sgn(t - y(x)) p(x,t) dt \tag{22}$$

$$= \int_{y(x)}^{\infty} q|t - y(x)|^{q-1} p(x,t) dt - \int_{-\infty}^{y(x)} q|t - y(x)|^{q-1} p(x,t) dt \tag{23}$$

$$\Rightarrow \int_{-\infty}^{y(x)} |t - y(x)|^{q-1} p(x,t) dt = \int_{y(x)}^{\infty} |t - y(x)|^{q-1} p(x,t) dt \tag{24}$$

For $q = 1$, we see that $y(x)$ is the conditional median of t.

$$\int_{-\infty}^{y(x)} p(x,t) dt = \int_{y(x)}^{\infty} p(x,t) dt \tag{25}$$

Finally, as $q \to 0$, the $y(x)$ given by the Minkowski loss is the conditional mode of t.

3 Information Theory

We want a function I which measures how much information you learn from observing some event E. We want it to satisfy some properties, mainly:

1. Highly probable E have low $I(E)$ and conversely $\rightarrow$ *rare events give more information.*
2. $I(E) \geq 0 \rightarrow$ *Information is non-negative.*
3. if $p(E) = 1$ then $I(E) = 0 \rightarrow$ *Events that always occur provide no information.*
4. If E_1, E_2 are independent events then $I(E_1 \cap E_2) = I(E_1) + I(E_2) \rightarrow$ *information due to independent events are additive.*

From 1. and 3. we see that I should be a function of the probability of an events occurence, i.e. $I(E) = f(p(E))$ for some f. From 4., given independent events E_1, E_2, we have that:

$$f(p(E_1)p(E_2)) = f(p(E_1 \cap E_2)) = f(p(E_1)) + f(p(E_2)) \tag{26}$$
$$f(x \cdot y) = f(x) + f(y) \tag{27}$$

If we assume that I is continuous, then only $I(E) = Klogp(E)$ satisfies (11) [6]. Finally, using 2., we see that $K < 0$. We can then define I as:

$$I(E) = -logp(E) \tag{28}$$

Where the choice of K decides the base of the logarithm. In this case we set it to 1 for clarity.

3.1 Entropy

We can extend this notion to a discrete Random Variable $X \sim p$ with finite domain $\mathcal{X}$. By defining the **Shannon Entropy** $H(X)$ as the average amount of information i.e.

$$H(X) = \mathbb{E}_{X\sim p}\{I(X)\} = \mathbb{E}_{X\sim p}\{-\log p(X)\} = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \tag{29}$$

We can also denote this as $H(p)$ where $p \sim X$ depending on what we want to emphasize. Note that WLOG we can assume that $p(x) > 0\ \forall x \in \mathcal{X}$. This is because we can use the convention that $0 \cdot log0 = 0$ (based on continuity arguments). Hence zero probability outcomes do not contribute to $H(X)$ anyways. We can further extend this for two Random Variables X, Y with finite domain $\mathcal{X} \times \mathcal{Y}$ by defining the **Joint Entropy** as:

$$H(X,Y) = -\sum_{x,y} p_{XY}(x,y)\log p_{XY}(x,y) \tag{30}$$

The **Conditional Entropy** is defined as:

$$H(X|Y) = \mathbb{E}_{X|Y}\{-\log p(X|Y)\} = -\sum_{x,y} p_{XY}(x,y)\log p_{X|Y}(x|y) \tag{31}$$

These quantities have nice properties:

1. *Non-negativity*: $H(X) \geq 0$, with equality only when X is a constant.

   PROOF: WLOG we assume that $p(x) > 0\ \forall x \in \mathcal{X}$. We have that $H(X) = -\sum_x p(x) log p(x) = \sum_x p(x) log p(x)^{-1} \geq 0$, since $p(x) > 0$ and $p(x)^{-1} \geq 1$. If $H(X) = 0$ then $\exists \alpha$ such that $p(\alpha)^{-1} = 1 \Rightarrow p(\alpha) = 1$. Hence X must be a constant, as needed.

2. *Chain Rule*: $H(X,Y) = H(X \mid Y) + H(Y) = H(Y \mid X) + H(X)$

3. *Monotonicity*: $H(X \mid Y) \leq H(X)$

3.2 KL Divergence

We can now look at the **KL Divergence** or **Relative Entropy**. This quantity measures the "distance" between two probability mass functions p and q.

$$KL(p||q) = \mathbb{E}_{X\sim p}\left\{\log \frac{p(X)}{q(X)}\right\} = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \tag{32}$$

The KL divergence has some nice properties.

1. $KL(p||q) \geq 0$ with equality iff $p = q$

   PROOF: If there exists $x \in \mathcal{X}$ such that $p(x) = 0$ and $q(x) > 0$, then $KL(p||q) = \infty$. Otherwise:

$$\begin{aligned} -KL(p||q) &= \mathbb{E}_{X\sim p}\left\{\log\frac{q(X)}{p(X)}\right\} \\ &\overset{(a)}{\leq} \log \mathbb{E}_{X\sim p}\left\{\frac{q(X)}{p(X)}\right\} \\ &= \log\sum_x p(x)\frac{q(x)}{p(x)} = \log\sum_x q(x) = 0 \end{aligned}$$

   Where (a) follows from Jensen's inequality. $KL(p||q) = 0$ only occurs when there is equality in Jensen's inequality, which only occurs when $p(x) = cq(x)$ for some c. Since $\sum_x cq(x) = c\sum_x q(x) = c \Rightarrow c = 1$, so $p = q$ as needed.

2. $KL(p||q)$ is strictly convex in each argument

3. $KL(p||q) \neq KL(q||p)$ so it is not a metric

4. We can decompose the KL divergence into two seperate terms:

$$KL(p||q) = \sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}} p(x)\log q(x) \tag{33}$$
$$= -H(p) + \mathbb{E}_{X\sim p}\{-\log q(x)\} \tag{34}$$
$$= -H(p) + CE(p,q) \tag{35}$$

Where the $H(p)$ is the Entropy and $CE(p,q)$ is called the **Cross Entropy**.

3.3 Mutual Information

We can quantify the amount of information obtained about one discrete random variable X, through another Y by defining the **Mutual Information** as:

$$I(X,Y) = \sum_{x,y} p_{X,Y}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \tag{36}$$

We again assume WLOG that $p(x,y) > 0\ \forall(x,y) \in \mathcal{X}\times\mathcal{Y}$. We note the following properties of I:

1. $I(X,X) = H(X) \rightarrow$ Sometimes the Entropy is called the **Self Information**
2. $I(X,Y) = KL(p_{XY}||p_Xp_Y)$
3. $I(X,Y) \geq 0$

   Proof. Notice that $I(X,Y) = KL(p_{XY}||p_Xp_Y) \geq 0$ by the positiveness of $KL(\cdot||\cdot)$ □

4. $I(X,Y) = H(p_X) + H(p_Y) - H(p_{XY})$

   Proof. We use property 2. of I and property 4. of $KL(\cdot||\cdot)$

$$\begin{aligned}
I(X,Y) &= KL(p_{XY}||p_Xp_Y)\\
&= -H(p_{XY}) + CE(p_{XY}, p_Xp_Y)\\
&= -H(p_{XY}) - \sum_{x,y} p_{X,Y}(x,y)\log p_X(x)p_Y(y)\\
&= -H(p_{XY}) - \left(\sum_{x,y} p_{X,Y}(x,y)\log p_X(x) + \sum_{x,y} p_{X,Y}(x,y)\log p_Y(y)\right)\\
&= -H(p_{XY}) - \left(\sum_{x} p_X(x)\log p_X(x) + \sum_{y} p_Y(y)\log p_Y(y)\right)\\
&= -H(p_{XY}) + H(p_X) + H(p_Y)
\end{aligned}$$

□

3.4 Differential Entropy

We can define the Entropy, KL divergence and Mutual Information for continuous random variables.

$$H(p) = -\int_{x\in\mathcal{X}} p(x)\log p(x)dx \tag{37}$$

$$KL(p,q) = \mathbb{E}_{X\sim p}\left\{\log\frac{p(X)}{q(X)}\right\} = \int p(x)\log\frac{p(x)}{q(x)}dx \tag{38}$$

$$I(X,Y) = \int\int p_{X,Y}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}dxdy \tag{39}$$

In the continuous case, the properties previously described hold except that the entropy is no longer necessarily non negative. For an example of this let $p = Uniform\left(\frac{1}{2},1\right)$. Then $H(p) = log(\frac{1}{2}) < 0$.

3.5 Entropy and Estimation

The KL divergence can be used within the Decision Theoretic framework. Semantically, $KL(p||q)$ represents how well some distribution q approximates the "true" p. Suppose we wanted to estimate a distribution p which we knew belonged to a Parametric Family $p \in \{p_\theta\}_{\theta\in\Theta}$. Let $\mathcal{A} = \{p_\theta\}_{\theta\in\Theta}$ and $\delta(D) = p_\theta$. Recall that this is similar to the Estimation problem described in 2.3.1 [2]. We can define the **Negative Log Loss**:

$$L(p,p_\theta) = -\log p_\theta(X) \tag{40}$$

This loss makes sense as $p_\theta(X)$ small means that the model has not taken into account X, and the corresponding loss will be large. The Cross Entropy is the corresponding risk function for this:

$$R(p,p_\theta) = \mathbb{E}_{X\sim p}\{-\log p_\theta(X)\} \tag{41}$$

This risk also makes sense. $KL(p,p_\theta) = -H(p) + L(p,p_\theta)$, and since $H(p)$ is constant, minimizing the KL is equivalent to minimizing the cross entropy. Since $KL(p,p_\theta) \geq 0$ we see that the minimum is attained at $L(p,p_\theta) = H(p)$, which occurs when $p_\theta = p$ i.e. when our prediction matches the "true" density.

[2] We modify the problem to make explicit the intention of estimating the density rather then the parameter. These goals are the same provided the parametric family is **identifiable**

3.5.1 Maximum Likelihood Estimation

We don't know p, so we cannot compute (41). Instead, we can use in its place the empirical density function $\hat{p}$, as defined in 1.1.1. Given X discrete, it turns out that the MLE for θ is the same as $\arg\min_{\theta\in\Theta} KL(\hat{p}||p_\theta)$. This is because:

$$\begin{aligned}
KL(\hat{p}||p_\theta) &= -H(\hat{p}) + CE(\hat{p}, p_\theta) \\
&= -H(\hat{p}) - \sum_{x\in\mathcal{X}} \hat{p}(x) \log p_\theta(x) \\
&= -H(\hat{p}) - \frac{1}{n} \sum_{x\in\mathcal{X}} \sum_{i=1}^{n} \delta(x, x^{(i)}) \log p_\theta(x) \\
&= -H(\hat{p}) - \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x^{(i)}) \\
&= -H(\hat{p}) - \frac{1}{n} l(\theta \mid x^{(1)}, \cdots, x^{(n)})
\end{aligned}$$

This provides a nice interpretation for the MLE - it is finding the $p \in \{p_\theta\}_{\theta\in\Theta}$ which minimizes the dissimilarity between the empirical distribution of the training set and itself as measured by the KL divergence. Conversly we can justify the use of the Cross Entropy loss through its equivalence to Maximum Likelihood.

On a final note, one may think that the quantity $KL(p_\theta||\hat{p})$ could be interesting. They would be wrong. This is since $p_\theta(x) = 0 \Rightarrow \hat{p}_\theta(x) = 0$ but $\hat{p}_\theta(x) = 0 \not\Rightarrow p_\theta(x) = 0$ since $\hat{p}_\theta(x) = 0$ only means that the particular value of x wasn't observed in the sample.

3.5.2 Maximum Entropy Principle

The **Principle of Maximum Entropy** (MaxENT) states that the probability distribution which best represents the "current state of knowledge" is the one with the largest entropy. More specifically, given some subset of distributions on $\mathcal{X}$ denoted as $\mathcal{M}$, we want to chose as our estimated distribution:

$$\arg\max_{q\in\mathcal{M}} H(q)$$

We may impose constraints to this in the form of **Testible Information**– statements about q with well-defined truth or falsity. The most basic of these is that $\int_\mathcal{X} q(x)dx = 1$. We now show a few maximum entropy distributions.

Theorem 3.1. *Let $X \sim p$ be a RV with finite support $\mathcal{X}$, $|\mathcal{X}| = k$, and $\mathcal{M} = \Delta_k$. The uniform density is the MaxENT density.*

Proof. We derive the following upper bound for $H(p)$

$$H(p) \leq \log k \tag{42}$$

To derive this inequality, let $q \sim Uniform$ on $\mathcal{X}$. We have that:

$$\begin{aligned} D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) \\ &= -H(p) + \sum_x p(x) \log k \\ &= -H(p) + \log k \end{aligned}$$

and so $H(p) = \log k - D(p||q) \Rightarrow H(p) \leq \log k$ as needed. Since $H(q) = \log k$ we can see that equality holds iff $p \sim Uniform$. □

So we have that, for densities with finite support and no testible information (apart from being a valid pmf), the MaxENT solution is uniform.

Theorem 3.2. *The MaxENT density for Random Variables $X_1 \in \mathcal{X}_1$ and $X_2 \in \mathcal{X}_2$ with $X_1 \sim p_1$ and $X_2 \sim p_2$ is $(X_1, X_2) \sim p_1 p_2$. i.e. higher entropy assumes independence.*

Proof. Properties 3. and 4. of I gives us that $I(X_1, X_2) \geq 0 \Rightarrow H(X_1) + H(X_2) \geq H(X_1, X_2)$, and so the maximal entropy of (X_1, X_2) is $H(X_1) + H(X_2)$. By definition this only occurs when $I(X_1, X_2) = 0$, which only occurs if $p_{1,2}(x_1, x_2) = p_1(x_1)p_2(x_2)\ \forall x_1, x_2 \in \mathcal{X}_1 \times \mathcal{X}_2$. □

Theorem 3.3. *The MaxENT of X with $\mathcal{X} = \mathbb{N}$ and with testible information $E(X) = \alpha$ is the Geometric Distribution $p(k) = \left(\frac{\alpha}{1+\alpha}\right)^k \frac{1}{1+\alpha}$*

Proof. We want to find the distribution which maximizes the entropy $H(p)$ satisfying the constraints $\mathbb{E}(X) = \alpha$ and $\sum_{i=0}^{\infty} p(i) = 1$. We form the Lagrangian:

$$L(p, \nu, C) = -H(p) + \nu \left(\sum_{i=0}^{\infty} ip(i) - \alpha \right) + C \left(\sum_{i=0}^{\infty} p(i) - 1 \right)$$

Taking the derivative w.r.t. $p(k)$ we get:

$$\frac{\partial}{\partial p(k)} L(p, \nu, C) = -\log p(k) - 1 + k\nu + C \tag{43}$$

$$\Rightarrow p(k) = \exp\{k\nu\} \exp\{C - 1\} \tag{44}$$

And using that $\sum_{i=0}^{\infty} p(i) = 1$ we have that

$$\sum_{i=0}^{\infty} \exp\{i\nu\} \exp\{C-1\} = 1 \Rightarrow \exp\{-C+1\} = \sum_{i=0}^{\infty} \exp\{i\nu\} \tag{45}$$

we substitute (45) into (43) to eliminate C

$$p(k) = \frac{\exp\{k\nu\}}{\sum_{i=0}^{\infty} \exp\{i\nu\}} \tag{46}$$

We then solve for α

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{\infty} \frac{k \exp\{k\nu\}}{\sum_{i=0}^{\infty} \exp\{i\nu\}} = \alpha \\
&\Rightarrow \sum_{k=0}^{\infty} k \exp\{k\nu\} = \alpha \sum_{i=0}^{\infty} \exp\{i\nu\} \\
&\overset{(a)}{\Rightarrow} \frac{\exp\{\nu\}}{(1-\exp\{\nu\})^2} = \frac{\alpha}{(1-\exp\{\nu\})} \\
&\Rightarrow \exp\{\nu\} = \frac{\alpha}{1+\alpha}
\end{aligned}$$

Where (a) comes from the geometric series. Finally, we sub this value into (46) to get the familiar formula:

$$p(k) = \left(\frac{\alpha}{1+\alpha}\right)^k \frac{1}{1+\alpha} \tag{47}$$

□

3.5.3 MaxENT and the Exponential Family

Let X be finite with $\mathcal{X}$ and k as defined before. Suppose we have feature functions $T_1(X), \cdots, T_d(X)$ and we define $\mathcal{M}$ as

$$\mathcal{M} = \Big\{ q : \underbrace{\mathbb{E}_q\{T_j(X)\}}_{\substack{model\ expected \\ feature\ count}} = \underbrace{\mathbb{E}_{\hat{p}}\{T_j(X)\}}_{\substack{empirical \\ feature\ count}} \ j = 1, \cdots, d \Big\} \tag{48}$$

Where our testible information are d *moment constraints*. Using the relation $H(p) = \log k - D(p||q)$ derived from theorem 3.1 we have the following alternative characterization of MaxENT:

$$\arg\max_{q \in \mathcal{M}} H(q) = \arg\min_{q \in \mathcal{M}} KL(q, Uniform) \tag{49}$$

We then pose the MaxENT problem as the following optimization problem:

$$\begin{aligned} \text{Minimize } & \sum_x q(x) \log \frac{q(x)}{u(x)} \\ \text{subject to } & q(x) \geq 0 \\ & \sum_x q(x) = 1 \\ & \sum_x q(x) T_j(x) = \alpha_j \ j = 1, \cdots, d \end{aligned}$$

References

[1] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[2] B. Vidakovic, "The likelihood principle." Slides.

[3] M. I. Jordan, "260 course notes." Slides.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[5] P. Hoff, "Bayes estimators." Notes, 2013.

[6] T. Carter, "An introduction to information theory and entropy." Slides, 2004.