Machine Learning Notes

Matthew Scicluna

December 13, 2017

1 Why Bayesian?

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.

1.1 Existence Of The Prior

We need a theoretical justification for why we assume the existence of a prior distribution on θ when doing Bayesian statistics. The theoretical justification requires the following assumption about the data. We say a sequence of random variables $x_1, \cdots$ are **Infinitely Exchangable** if for any n, and any permutation of size n $\pi_{1:n}$

$$p(x_1, \cdots, x_n) = p(x_{\pi_1}, \cdots, x_{\pi_n}) \tag{1}$$

The **De Finetti Theorem** says that a sequence is infinitely exchangible iff for any n,

$$p(x_1, \cdots x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) P(d\theta) \tag{2}$$

For some measure P on θ.
Note: if θ has a density then $P(d\theta) = p(\theta)d\theta$. What this theory means is that given the assumption of exchangable data (and iid $\Rightarrow$ exchangable) then there must exist a θ, $p(x|\theta)$ and distribution P on θ. Also note: θ may be infinite!

1.2 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 other 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.

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**

$$R(P, \delta) = \mathbb{E}_{D \sim P}\{L(P, \delta(D))\} \tag{3}$$

If we want to compare decision rules δ_1, δ_2 using (3), 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 (3). One way to take this into account is to use the **Minimax Criteria**: the optimal δ minimizes the Frequentist Risk in worst case scenerio.

$$\delta_{minimax} = \min_{\delta} \max_{P \in \mathcal{P}} \; R(P, \delta) \tag{4}$$

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 (3).

$$\delta_{bayes} = \arg\min_{\delta} \int_{\Theta} R(P_\theta, \delta) \pi(\theta) d\theta \tag{5}$$

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

$$R_B(\delta|D) = \int_{\Theta} L(P_\theta, \delta) p(\theta|D) d\theta \tag{6}$$

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).

$$\delta_{post} = \arg\min_{\delta} \; R_B(\delta|D) \tag{7}$$

Notice that in (6), 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 (6), meaning that (7) gives the undisputed optimal δ!

Note that the Frequentist can still use (7) by interpreting it as (5) 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 [4] 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 some cost, for example: $L(\theta, \delta(D)) = ||\theta - \delta(D)||^2$. 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 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 the loss over all possible datasets, we can compare how much of the loss is due to the Bias and how much to the Variance. 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.

2.3.2 Prediction

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

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:

$$L(P, f) = \mathbb{E}_{(X,Y)\sim P_{XY}}\{l(Y, f(X))\} \tag{8}$$

This is often called the Risk in Machine Learning. Note that we do not know (8), but we can approximate it using the below formula. This is called **Empirical Risk Minimization**

$$L(P, f) = \frac{1}{n}\sum_{i=1}^{n} l\left(Y^{(i)}, f(X^{(i)})\right) \tag{9}$$

2.4 Loss Functions

Given $\mathcal{A} = \Theta$ and $L(\theta, a) = ||\theta - a||^2$ we have that $\delta_{post}(D) = \mathbb{E}\{\theta|D\}$. This is because:

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

and taking the derivative and setting to 0 yields:

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

The loss $L(\theta, a) = ||\theta - a||^2$ is so popular that it has its own name – the squared loss. It falls in a family of loss functions called the **Minkowski Loss**. Some examples in this family include:

1. $L(P_\theta, a) = ||\theta - \delta(D)||^2$
2. $L(P_\theta, a) = ||\theta - \delta(D)||^1$
3. $L(P_\theta, a) = \lim_{p\to 0} ||\theta - \delta(D)||^p$

These distributions result in the mean, median and mode of the data, respectively.

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:

$$f(P(E_1)P(E_2)) = f(P(E_1 \cap E_2)) = f(P(E_1)) + f(P(E_2)) \tag{10}$$

$$f(x \cdot y) = f(x) + f(y) \tag{11}$$

If we assume that I is continuous, then only $I(E) = KlogP(E)$ satisfies (11) [5]. Finally, using 2., we see that $K < 0$. We can then define I as:

$$I(E) = -logP(E) \tag{12}$$

Where the choice of K decides the base of the logarithm.

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.

$$H(X) = \mathbb{E}_{X\sim P}\{I(X)\} = \mathbb{E}_{X\sim P}\{-\log P(X)\} = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \tag{13}$$

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:

$$H(X,Y) = -\sum_{x,y} p_{XY}(x,y)\log p_{XY}(x,y) \tag{14}$$

The **Conditional Entropy** is defined as:

$$H(X|Y) = \mathbb{E}_{X|Y}\{-\log p(X|Y)\} = -\sum_{x,y} p_{XY}(x,y)\log p_{X|Y}(x|y) \tag{15}$$

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 $H(X) = -\sum_x p(x)logp(x) = \sum_x p(x)logp(x)^{-1} \geq 0$, 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.

$$KL(p||q) = \mathbb{E}_{X\sim p}\left\{\log\frac{p(X)}{q(X)}\right\} = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \tag{16}$$

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:

$$KL(p||q) = \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} p(x) \log q(x) \tag{17}$$

$$= -H(p) + \mathbb{E}_{X \sim p}\{-\log q(x)\} \tag{18}$$

$$= -H(p) + CE(p, q) \tag{19}$$

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

The KL divergence can be used within the Decision Theoretic framework. We elaborate in the following subsections.

3.2.1 Cross Entropy as a Loss Function

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 $p \in \{p_\theta\}_{\theta \in \Theta}$. We can put this into our decision theoretic framework. Let $\mathcal{A} = \{p_\theta\}_{\theta \in \Theta}$ and $\delta(D) = p_\theta$. We can use the Cross Entropy as a loss function.

$$L(p, p_\theta) = \mathbb{E}_{X \sim p}\{-\log p_\theta(X)\} \tag{20}$$

This is because $KL(p, p_\theta) = -H(p) + L(p, p_\theta)$, 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.

3.2.2 Maximum Likelihood Estimation

We don't know p, so we cannot compute (20). Instead, we can use in its place the empirical distribution $\hat{p}$, i.e.

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x, x^{(i)}) \tag{21}$$

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. Because of this, the Cross Entropy loss is also known as the **Negative Log Loss**.

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.

3.3 Differential Entropy

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

$$
H(p) = -\int_{x \in \mathcal{X}} p(x) \log p(x) d\mu(x) \tag{22}
$$

$$
KL(p, q) = \mathbb{E}_{X \sim p} \left\{ \log \frac{p(X)}{q(X)} \right\} = \int_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x) \tag{23}
$$

In the continuous case, the properties previously described hold except that the entropy is no longer necessarily non negative.

3.4 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:

$$
I(X, Y) = \sum_{x,y} p_{X,Y}(x, y) \log \frac{p_{XY}(x, y)}{p_X(x) p_Y(y)} \tag{24}
$$

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. $I(X,Y) = KL(p_{XY}||p_X p_Y)$
3. $I(X,Y) \geq 0$

   PROOF: Notice that $I(X,Y) = KL(p_{XY}||p_X p_Y) \geq 0$ by the positiveness of $KL(\cdot||\cdot)$.
4. $I(X,Y) = H(p_X) + H(p_Y) - H(p_{XY})$

   PROOF:

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

3.5 Maximum Entropy Principle

3.5.1 Examples

We can show that for discrete spaces, the uniform density is the Maximum Entropy. Let $X \sim p$ and $Dom(X) = \mathcal{X}$ with k elements and $q \sim Unif$ on $\mathcal{X}$. Then $D(p||q) = -H(X) + logk$

PROOF:

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

An upper bound for $H(X)$ is $logk$ since $H(X) = logk - D(p||q) \Rightarrow H(X) \leq logk$.

The maximum entropy solution for Random Variables is Independence From the previous result we have 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 $p_{1,2}(x_1, x_2) =$

$p_1(x_1)p_2(x_2)$ $\forall x_1, x_2 \in \mathcal{X}_1 \times \mathcal{X}_2$. This can be seen directly from the definition of I and using the strict positivity of $p(x_1, x_2)$.

The maximum entropy solution for continuous random variables with given mean and variance is Normal.

4 Notes

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.

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

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