Machine Learning Notes

Matthew Scicluna

December 9, 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.

2.2.3 Examples

An example of a Bayes Action is: 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:

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

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.*

It turns out that if we assume that I is continuous, only one function satisfies the above [5]

$$I(E) = -logP(E) \tag{8}$$

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.

$$H(X) = \mathbb{E}_{X\sim P}\{I(X)\} = \mathbb{E}_{X\sim P}\{-\log P(X)\} = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \tag{9}$$

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:

$$H(X,Y) = -\sum_{x,y} P_{XY}(x,y)\log P_{XY}(x,y) \tag{10}$$

The **Conditional Entropy** is defined as:

$$H(X|Y) = \mathbb{E}_{X|Y}\{-\log P(X|Y)\} = -\sum_{x,y} P_{XY}(x,y)\log P_{X|Y}(x|y) \tag{11}$$

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*: $H(X \mid Y) \leq H(X)$

We can extend these Entropy definitions to the continuous case, where the above properties hold - except for non negativity.

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.

$$KL(p||q) = \mathbb{E}_{X\sim p}\left\{\log\frac{p(X)}{q(X)}\right\} = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \tag{12}$$

The KL divergence has some nice properties.

1. $KL(p||q) \geq 0$

   PROOF: Use that

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

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

We can decompose the KL divergence into two seperate terms:

$$\begin{aligned} KL(p||q) &= \sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}} p(x)\log q(x) \\ &= -H(p) + \mathbb{E}_{X\sim p}\{-\log q(x)\} \end{aligned}$$

Where the first term is the Entropy and the second term is called the **Cross Entropy**. The Cross Entropy makes a good Loss function.

3.2.1 Maximum Likelihood Estimation

Given a Parametric Family $\{P_\theta\}_{\theta\in\Theta}$ it turns out that the MLE for θ is the same as $\arg\min_{\theta\in\Theta} KL(\hat{p}, p_\theta)$, where $\hat{p}$ is the empirical distribution, i.e.

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x, x^{(i)}) \tag{13}$$

This is since:

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

3.3 Mutual Information

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

$$I(X,Y) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y) log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \tag{14}$$

We again assume WLOG that $p(x_1, x_2) > 0\ \forall (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$

1. CLAIM: $I(X_1, X_2) \geq 0$

   PROOF: Notice that $I(X_1, X_2) = D(p_{1,2}||p_1p_2) \geq 0$ by the positiveness of $D(\cdot||\cdot)$.

2. We want to express $I(X_1, X_2)$ as a function of $H(X_1), H(X_2)$ and $H(X_1, X_2)$.

$$\begin{aligned}
I(X_1, X_2) &= \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1,x_2) log \frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)} \\
&= \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1,x_2) log p_{1,2}(x_1,x_2) - p_{1,2}(x_1,x_2) log p_1(x_1)p_2(x_2) \\
&= -H(X_1,X_2) - \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} \Big(p_1(x_1)p_2(x_2) log p_1(x_1) - p_1(x_1)p_2(x_2) log p_2(x_2) \Big) \\
&= -H(X_1,X_2) - \sum_{j=1}^{2} \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_1(x_1)p_2(x_2) log p_j(x_j) \\
&= -H(X_1,X_2) - \sum_{j=1}^{2} \sum_{x_j\in\mathcal{X}_j} p_j(x_j) log p_j(x_j) \\
&= -H(X_1,X_2) + H(X_1) + H(X_2)
\end{aligned}$$

   And so we can represent $I(X_1, X_2)$ using $H(X_1), H(X_2)$ and $H(X_1, X_2)$, as needed.

3. 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 $p(x_1, x_2)$.

Mutual Information $I(X, X) = H(X)$

3.4 Maximum Entropy Principle

Density Estimation - Given a Parametric Family $\{P_\theta\}_{\theta\in\Theta}$ we would like to use data X to choose a P from our Family.

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:

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

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

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.