CMPUT-267: Thought Questions

Joshua J.George

September 16, 2021

Thought questions for chapters 1-3

Q1 (sec 2.1): Given co-domain of any function in the probability space is $[0, 1]$.This makes complete sense as if any event is more likely to happen it is closer to 1 and 0 otherwise. I was wondering about *negative probability.* My question is, are negative probabilities existent in the day-to-day world?(excluding quantum mechanics). There's a very satisfactory proof given online disproving this- If you suppose just one subset E_k to have negative measure, then consider $E_k \cup E_l$. Clearly $E_l \subset E_k \cup E_l$ but the increasing characteristic of probability is lost as $p\left(E_l\right) > p\left(E_l \cup E_k\right)$, but according to a paper by **Richard Feynman** he describes a situation in which a man has a certain number of apples say 5 and later 10 apples are taken away from him (impossible realistically) and 8 given back to him. In the real situation there must be special limitations of the time in which the various apples are received and given since he never really has a negative number, yet the use of negative numbers as an abstract calculation permits us freedom to do our mathematical calculations in any order simplifying the analysis enormously. Does this translate to probabilities as well?

Q2 (general topic): Till now in the course we have been discussing the about statistics in the real field. In $\mathbb{R}$ everything is plotted/measured/evaluated realistically. (we can measure it realistically). However in $\mathbb{C}$ how do we plot events with i on the yaxis. How do we define statistics/probability in the Complex field? (or is it imaginary)

Thought questions for chapters 4-6

Q3 (sec 5.1-5.2): We know that the MLE is a distributional parameter that maximizes the probability of observing a particular set of data (maximum of some likelihood function) and MAP gives you the value which maximises the posterior probability.MAP is essentially the same as MLE but with a uniform prior.

MLE: $f_{\mathrm{MLE}} = \underset{f \in \mathcal{F}}{\operatorname{argmax}}\ p(\mathcal{D} \mid f)$
MAP:$f_{\mathrm{MAP}} = \underset{f \in \mathcal{F}}{\operatorname{argmax}}\ p(\mathcal{D} \mid f) p(f) = \underset{f \in \mathcal{F}}{\operatorname{argmax}}\ p(f \mid \mathcal{D})$

Claim: MAP estimate eventually converges to MLE with a uniform prior and as $n \to \infty$ (not sure if this is an acceptable proof for this, the gradient method is clean, just wanted to try something new).

Proof: By the definition of MAP,

MAP $:= f_{\text{MAP}} = \underset{f \in \mathcal{F}}{\operatorname{argmax}}\, p(\mathcal{D} \mid f)p(f) = \underset{f \in \mathcal{F}}{\operatorname{argmax}} \ln p(\mathcal{D} \mid f) + \ln p(f)$
Now by assumption, $p(f)$ is uniform **everywhere** and $p(f) := (\frac{1}{m})^n$ where m is a constant. Therefore, as $n \to \infty, p(f) \to c \approx 0$, where c is a constant. We get,

$\underset{f \in \mathcal{F}}{\operatorname{argmax}} \ln p(\mathcal{D} \mid f) + \ln p(f) = \underset{f \in \mathcal{F}}{\operatorname{argmax}} \ln p(\mathcal{D} \mid f) + \text{constant} = \underset{f \in \mathcal{F}}{\operatorname{argmax}}\, p(\mathcal{D} \mid f) = f_{\text{MLE}}$

Here we assumed $p(f)$ is uniform everywhere or constant. I was wondering if it is possible for MAP to eventually converge to MLE without a uniform prior or **almost everywhere** uniform distribution?

Q4(sec 4.2):We know that the first and second derivative test tells us about whether a point is a local max or min. What about the third derivative?
Does the third derivative tell us anything about any arbitrary continuous function? Some sources say "the rate of change of curviness ie concave down to concave up implies positive derivative". Is this true for even a discrete function curved between disjoint intervals? (here I am assuming the discrete derivative of a function $f(n)$, denoted $\Delta_n f(n)$, to be defined as$f(n+1) - f(n)$)

Thought questions for chapters 10-12

Q5(sec 10.1):So we have defined Lasso Regressor in the class as-

$$c(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{x}_i^\top \mathbf{w} - y_i\right)^2 + \lambda \sum_{j=1}^{d} |w_j|$$

and we stated the following "Instead, we use gradient descent to compute a solution to $\mathbf{w}$. The ℓ_1 regularizer, however, is non-differentiable at 0. Understanding how to optimize this objective requires a bit more optimization background; we leave it for a future course.: I was wondering if this is a satisfactory proof for showing existence of a solution?(we know the function is not differentiable but if we can optimize the function there has to be a way to solve it and find the best solution)
Proof: If I can show the lasso is convex then there exists a global minimum. Expand $\left(\mathbf{x}_i^\top \mathbf{w} - y_i\right)^2$

$$f := w.\mathbf{x}_i^\top.\mathbf{x}_i.w^\top - 2.\mathbf{x}_i^\top.w + y_i.y^\top$$

Well let $\mathbf{x}_i^\top.\mathbf{x}_i$ to be positive definite then it implies f is convex (as sum of convex functions). Consider taylors theorem :

$$f(x+\xi) = f(x) + (\mathrm{grad}\, f)(x)\cdot\xi + \frac{1}{2}(\mathrm{Hess} f)(x+\theta\xi)\xi\cdot\xi$$

which is equivalent to

$$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(x+t(y-x))(y-x)$$

for $t \in [0,1]$ Since $\nabla^2 f$ is everywhere positive semi-definite, the quadratic term in this equation is always non-negative. Thus

$$f(y) \geq f(x) + \nabla f(x)^T(y-x)$$

implies its convex. Therefore f is convex. Also we know $\lambda\sum_{j=1}^{d}|w_j|$ is convex. Therefore the Lasso is convex as sum of convex functions $\implies$ it has a global minimum. My second question under this topic is why is this so- "the ℓ_1 regularizer penalizes large values in $\mathbf{w}$. However, it also produces more sparse solutions, where entries in $\mathbf{w}$ are zero." ?

Q6(sec 10.3)I came across this paradox while reading stuff about Bias Variance tradeoff- Stein's paradox: A case of multivariate Bias Variance decomposition. Suppose ρ is a vector of three components and independent of each other and normally distributed with unknown mean. For the 1 Dimensional case, if we take a single sample from the distribution we guess that the sample is the mean. For the 2-D case,we have a 2-D Gaussian as well, here the mean would be a 2-D vector and equal to a random single sample take. What about the 3-D case?- How is this any different from the 1-D and 2-D case? Well the paradox tells us that for say the 3-D case if we shrink the estimator towards any arbitrary point then we get a biased estimator and correlated but with a low MSE as we consider three components with independent distribution but a good correlated estimator. How is this possible?