08:00UTC Session on Metaverification

Keith Mitchell - Outcome-conditioned decompositions of proper scores

Discussion:

Deryn: We use Brier Score and its decomposition; looking forward to reading your paper. Your example showed that with conditioning on the forecast you're able to discern quite different properties between the two forecasts; is it in the decimal points, it's there in the original decomposition but it's hard to distinguish because it's in the decimal places and the second decomposition makes it a bit more obvious or is it just that you're looking at it in a completely different way?

KeithMitchell: Are you referring to the motivation slide? If you express them as fractions, these values are exact values, no difference between the forecast decompositions even in the decimal place. Admit this is artificial example, but demonstrates that even for exact correspondence between the 2 forecasters in terms of forecast conditioned decomposition we are able to see evident differences in the terms of the OCD so if we were to express these values not as decimals, but as fractions it would display exact correspondence in FCD.

DerynGriffiths: Users want to know decomposition by observation, but you only know that after the even, but if its telling you about the different properties of the two forecasts that's very interesting.

RobertTaggart: How do you find the dual scoring function? Can you give an example?

KeithMitchell: Example: last slide in the presentation. BS rule replace outcome with average forecast under the distribution of the forecast. See the last slide for details. There is as yet no characteristic theorem similar to Savage's theorem for ordinary scoring rules which gives us a functional form for R* given S, which would be a nice result. It does involve looking for a scoring rule which matches the qualities required for a dual scoring rule. Haven't managed to find this for all scoring rules, yet. A characteristic theorem would be nice.

RogerHarbord: What other scoring rules have you investigated? Have you derived outcome-condidtioned decomposition for the CRPS?

KeithMitchell: Yes, we have. Have done the rank probability scores in general, Brier Scoring rule, Ignorance scoring rule and we've looked at things like alpha quantile and predicted interval scoring rule. Have looked at a few, but not all. There is scope for working a bit more on this type of problem. Working towards a characteristic theorem, but are not there yet.

Martin Leutbecher – Understanding the link between ensemble mean error variance, spread-error ratio, mean error and the CRPS

Discussion:

[08:39] DerynGriffiths: Do you think a similar technique could be used for assessing rainfall or wind speed? (Liked by RogerHarbord)

[09:10] MartinLeutbecher: Deryn (08:39): the technique can be applied to any scalar variable. We expect that the assumption of normality will result in a poorer approximation of the CRPS for variables like rainfall and to a lesser extent wind speed. As said earlier, it may be possible to extend this work to other families of distributions.

[08:45] MarionMittermaier: I guess that is because our real forecast systems have inherent biases, which we want to minimise. Our systems are "naturally" biased. (liked by ManfredDorninger)

BethEbert: You showed the decomposition into the 3 terms. Have you had a chance to explore this with actual model results?

MartinLeutbecher: Yes, in the paper at the end of the references we discuss a case where we try and understand the degradation in a particular model change in the temperature in the tropics. The change in the approximated CRPS is dominated by the bias term; the error of the ensemble mean is a minor term and the spread-error ratio has a neglible contribution in this change. Plan is to implement in routine verification so developers can see what is responsible for decreases and increases in CRPS.

MarionMittermaier: Post-processing. Your conclusions state that its medium-range upper air variables. I guess the assumption is that it won't hold if you're looking at surface variables at shorter range? Is that how you would interpret that?

MartinLeutbecher: I don't necessarily think that. It's just that we focussed on one application. Want to leave that to others to find out. For some variables, like precipitation, the assumption of Gaussianity will be a problem; too drastic. Would like to mention that you could extend the methodology for different distributions – end up with a different expected value of the CRPS.

MarionMittermaier: Yes, if you could come up with an analytical expression.

ChrisFerro: What are the implications of your final point for how weather forecasting models should be tuned, should we not be tuning to proper scoring rules?

MartinLeutbecher: May have to go away from optimising the CRPS of raw model output. May want to optimise the CRPS after bias correction, potentially. Would still apply a proper scoring rule, but would apply bias correction before that. In puristic terms does that mean not using proper scoring rule; there is some choice to be made how you correct biases which could influence what you do. For developing model uncertainty schemes we have found we converge to the wrong answer if we optimise for minimum CRPS of raw model output.

[08:49] ChiaraMarsigli: Thank you very much, Martin, very interesting work. The natural next question for me after the one of C. Ferro would then be: and how do we perform the bias correction? Have you ever found a convincing way of doing it, in presence of its weather dependence? (Liked by MarionMittermaier and DerynGriffiths)

[09:17] MartinLeutbecher: Chiara (question at 08:49): Regarding how to bias correct in the context of this work, I think that simple approaches that respect geographical and seasonal variability would be a good starting point.

RogerHarbord: Looking at performance of our post-processing system, have been looking at whether ensemble spread is appropriate we've been looking whether ensemble spread matches s.d. of ensemble mean rather than mean square error of the ensemble mean? Does this fit in with your last point?

MartinLeutbecher: Yes. That is the right thing to do. At ECMWF, senior management look at CRPS of raw model output. If that's degraded we have a lot of explaining to do.

ParomitaChakroborty: Would like to know more about the slide where you compare the CRPS with full gaussian distribution. What is the difference?

MartinLeutbecher: On RHS we have use exactly the formula based on epsilon, b and sigma from the actual NWP data. On LHS it's the actual CRPS computed in the standard way.

Alexander Jordan -Evaluating probabilistic classifiers: Reliability diagrams and score decompositions revisited

[09:15] Ashrit, Raghavendra: Do you use the sample climatology??

[09:20] Jordan,Alexander: Ashrit, Raghavendra (09:15): In the score decomposition, we use the sample climatology as the reference forecast.

[09:16] Ashrit, Raghavendra: How does change in bin affect?? clim line

RogerHarbord: Our bins are often small in number. We have ensemble forecasts with small numbers of members, so have naturally got small numbers of bins fixed by the ensemble size, is it still worth applying these ideas?

AlexanderJordan: Boils down to whether the bins are populated well enough, which I guess in this case they are.

RogerHarbord: Less so, if you take extreme thresholds. We see jumpy behaviour at extreme thresholds.

AlexanderJordan: You will get a more stable behaviour there. This approach works equally well for discrete and continuous forecast values. Could also say there's no need to do equidistant binning, or in this case for every discrete forecast value. Might as well use quantile-based binning. Everything is not equidistant, but equally populated which works well for continuous forecasts when you have unique forecast values, but you will run into the same problem of having to make decisions when the same value occurs multiple times. Say you want to look at the 90 percent value. When you have too many values, do you assign to left, or right bin? How do I assign observed values to left or right bin? Yes, using this approach would give more stability in edge cases, but in your case quantile-based binning is potentially finnicky as well. So, it could be useful...

DerynG: I have read the pre-print already, the animation slide of the pooling technique was useful. I would like to be using the technique. In paper you've put an area round the diagonal where you expect the line to fall if you have a reliable forecast. Do you have any advice for displaying 2 forecast sources on one graph? Do you have a nice visualisation?

AlexanderJ: Default choice in package is that there is no display of uncertainty; no consistency bands, no confidence bands and no histograms. Simply get the lines.

DerrynG: Confidence intervals are the important bit! May have to display 2 side-by-side then. Might have to explore that.

AlexanderJ: If there's only 2 you could get away with shading. Works better if they are both around the estimate than if they are both around the diagonal.

MichaelFoley: On traditional reliability diagrams you join the dots to connect between bins and to give a sense of what you might expect in the cracks between where you've binned. Do you do an equivalent for this method?

AlexanderJordan: [see slide 14] We've made the constant pieces thicker than the connections; there's no question about the estimate there; isotonic regression says these pieces have to be constant so they're more pronounced. The interpolation between constant pieces, anything would be allowed that maintains isotonicity. This is the default choice for continuous forecast values (see slide for illustration). If you have discrete forecast values you'd get points.

MichaelFoley: So, if I had a forecast system giving those forecast values, intuitively you'd expect if it went from 0-25 – 0.3 prob the expectation probability would really go up if you had enough samples, rather than being constant, based on the data you actually have?

AlexanderJ: You mean enforce a strict monotonicity? If you were to do that, the theorem I've stated won't hold anymore. You could have cases where the component becomes negative again. Under the isotonicity assumption, this is the only solution that minimises the loss for every proper scoring rule. So, if you pick anything else that is isotonically increasing that will have at least the same score or higher. If I enforce strict monotonicity then it would probably be higher. So, I guess you can but then you lose some of the theoretical properties.

[09:15] DerynGriffiths: Michael Foley, I will take a horizontal line in the reliability diagram over the noisy up and down any day for our use of it. (Liked by MichaelFoley)

IanJoliffe: Just a couple of comments. Similar ideas can used when doing choice of bins in goodness of fit tests. It can be surprising how much results can change if you tweak the bins a bit in a goodness of fit test. Monotonicity is a nice idea, deviating from it is undesirable. On the other hand when there is jumpiness, it might be nice to know about it. There could be information in there that you otherwise might miss. It might be nice to flag these inconsistencies and investigate whether there's something which explains them.

AlexanderJordan: You get some indication by long constant pieces across and you will still see the miscalibration component which is automatically shown in the software package when you print the results. Are you worried about the isotonicity assumption in general?

IanJoliffe: No, it seems a sensible thing to do.

Sebastian Lerch: Evaluating probabilistic forecasts with scoring rules

RobertTaggart: Just putting my interest in the generalised gamma distribution for the CRPS and also for weighted CRPS.

SebastianLerch: We have been discussing weighted CRPS, need to discuss how to best to implement this, may be done only for sample-based forecasts. Terms of weighting function which can be used need to be chosen.

MichaelFoley: We do a lot of work in python. Can one import R packages in a python context or does that get messy?

SebastianLerch: We've experimented with this a bit; it's possible to call R packages from python. Technically not difficult. Will add some computational overhead. Evaluation will be fine. But if you want to use the internal score computation this may not be feasible. Ideal would be to have a package with a core written in C, for example and then have APIs in python and R but this would add a lot of work maintaining it.

[09:40] MarionMittermaier: Rpy package.... our experience with this has been that it is very slow.... i.e. using python to handle I/O and use R to compute stats... we found it prohibitive for ensemble applications. (Liked by MichaelFoley)

ParomitaChakroborty: Calculating CRPS considering some thresholds? Or without? Which is better? If we consider mean CRPS will it still be a proper score?

SebastianLerch: Regarding the second bit of the question, with the mean CRPS you are estimating expected CRPS which is still a proper scoring rule. In terms of thresholds, this will relate to weighted scoring rules. Some of the talks have focussed on score decomposition, which hasn't been included in the package. In terms of thresholds, this is possible if you use thresholds inside of the CRPS in terms of the weighting function. It's not in the package now, but will be considered for the future.

[09:45] DerynGriffiths: Chakraborty,Paromita I think the CRPS and the thresholded CRPS are measuring different things. For the things they are measuring they are each proper. Maybe someone else can comment on this.

[09:47] RobertTaggart: Gneiting/Ranjan 2011 is a good reference for weighted or thresholded CRPS

MartinLeutbecher: Can you explain why the log-score computed with a kernel density estimate is fragile? Too small?

SebastianLerch: [see slide 13, titled Simulated Forecast Distrubutions]. All details can be found in the paper by Krüger, F., Lerch, S., Thorarinsdottir, T.L. and Gneiting, T. (2020). Quick explanation we have proposed a notion of consistency which formalises the notion that if your ensemble size, or sample size goes to infinity then the estimated forecast distribution should result in the same expected score as the true forecast distribution. It is possible to show that the CRPS will be consistent under very minimal assumptions. Log score will require much more strict assumptions, particularly if you don't have an independent sample.

RogerHarbord: A comment about Martin's thought about the problems with estimating ignorance score from a sample. I've been wondering if it would somehow be possible to use climatology as a Bayesian prior to give a more stable estimate of the ignorance score

SebastianLerch: Could use a parametric estimate instead of a kernel density estimation. This makes the estimation more stable, will assume that the forecast is from this parametric distribution. I think this can be reasonably assumed for example, looking at climatology would be a valuable way to go.

[09:47] MartinLeutbecher: Harbord, Roger, this would be a bit like scoring a post-processed ensemble?

[09:50] RogerHarbord: MartinLeutbecher Yes, i guess it would be... I haven't thought it through!

[09:50] Ferro, Chris: Regarding Sebastian's comment about computing the log score after fitting a parametric model, Stefan Siegert's paper is relevant: https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.3447

BethEbert: I've got one for Alexander with the PAV which I thought was a nifty approach particularly for small sample sizes. How sensitive is this to the removal of samples? Is it pretty stable? For example, if you had 30 and you removed 2 of them, are your results liable to be quite similar?

AlexanderJordan: Depends on the location of forecast values; if it's centred it probably won't change much. On the boundaries everything is unstable it's still not great with the PAV. We don't get huge spikes, but on lowest/highest forecast values these tend to be 0 or 1.

ChiaraMarsigli: Started from Martin's presentation. We are struggling to understand if we need more spread in our ensembles. We see that its under-dispersive, but we aren't sure if it's really under-dispersive because of representativeness and presence of bias. Would be good to have a strategy for this. Framework is clear - need to take care of observational uncertainty and take care of the bias. Agree with Marion's comment (see below). This is one of the key issues for ensemble development now.

[08:57] MarionMittermaier: I guess what it also shows is that the bias remains one of the key "issues" for verification metrics (and model development!). I am reminded of the Extremal Dependency Score paper which specifically also required that forecasts are debiased before the score is applied. This was for deterministic forecasts in the context of extremes. Different scenario, but still the same bias causing a problem, which needs to be eliminated in order to get a clear signal from the score. It is therefore crucial that we understand/quantify the bias. Maybe it is the FIRST thing we should be looking at, before we look at anything else? Second is knowing what impact that bias may have on the metrics we compute and whether the results can be trusted, even when a proper score is used. We need to account for account for it, especially if we're not aware of it or it can have harmful effects.

DerynGriffiths: I can see a lot of value in doing bias correction before doing verification, but I think for rare events, for example EDS, bias correction is done on too small a sample. I think it's going too far unless you have a big enough sample.

MarionMittermaier: Can get into trouble thinking empirically here. This is where taking more parametric approaches and using extreme value theory might help here.

RobertTaggart: Question for Alexander. I have a really perverse case for reliability diagrams. Suppose you have a really skilful forecaster who whenever they thought there was a high chance of event occurring, they forecast the complementary probability, a low probability and vice versa. What does assumption of isotonicity look like on the graph? On a typical reliability diagram the reliability line will go from the top left to bottom right – won't get that with isotonicity. Do the confidence intervals give you an indication that the perverse case is coming up?

AlexanderJordan: It will be constant at the marginal frequency. They will all get merged into a single bin. Indications will be the long constant bin, and you would see that in the missed calibration component which would be almost off the chart.

ChiaraMarsigli: Back to previous point of DerynGriffiths. Key issue is the sample. Stratify to isolate the bias relevant to that situation, then the sample size is small. Has anyone tried to use this bias removal effectively, after accounting for conditional verification and to see if the spread-skill relationship is more clear? Does anyone have experience of this?

[Nobody admitted to having tried this!]