Proper scoring rules

The coursework will be available from the start of lab 4, tomorrow.
No lecture in week 5.
There will be a few tutors available in the week 5 lab session to answer general questions.
A half-way feedback form will be available in Lab 4.

Some essential and useful probability theory

Expectation and variance

$$\mathsf{E}_F[h(Y)] = \sum_{k \in K} h(k)\, f_Y(k), \quad Y \sim F, \text{ prob. fcn } f_Y(\cdot) = \mathsf{P}_F(Y = k), \text{ discrete outcomes } K$$

$$\mathsf{E}_F[h(Y)] = \int_D h(y) f_Y(y)\, \mathrm{d}y, \quad Y \sim F, \text{ prob. density } f_Y(\cdot), \text{ continuous outcomes } D$$

$$\mathsf{Var}_F(Y) = \mathsf{E}_F\left\{[Y - \mathsf{E}_F(Y)]^2\right\} = \mathsf{E}_F(Y^2) - \mathsf{E}_F(Y)^2$$

Example: $Y \sim \mathsf{N}(\mu, \sigma^2)$, and we know $\mathsf{E}(\mathrm{e}^Y) = \exp(\mu + \sigma^2/2)$. What is $\mathsf{Var}(\mathrm{e}^Y)$?

Some essential and useful probability theory

Law of total expectation, or "The tower property"

$$
\begin{aligned}
\mathsf{E}(Y) &= \mathsf{E}\left[\mathsf{E}(Y \mid X)\right] \\
\mathsf{Var}(Y) &= \mathsf{E}\left[\mathsf{Var}(Y \mid X)\right] + \mathsf{Var}\left[\mathsf{E}(Y \mid X)\right]
\end{aligned}
$$

Example: $\mu \sim \mathsf{N}(5, 4)$, $(Y \mid \mu) \sim \mathsf{N}(\mu, 1)$. What is $\mathsf{E}(Y)$ and $\mathsf{Var}(Y)$?

Scores

- We want to quantify how well our predictions represent the test data.
- We define *scores* $S(F, y)$ that in some way measure how well the prediction F matched the actual value, y.
- The scores defined here are *negatively oriented*, meaning that the *lower the score, the better*.

Squared errors and log-likelihood scores

- Squared Error (SE): $S_{\text{SE}}(F, y) = (y - \widehat{y}_F)^2$,
  where $\widehat{y}_F$ is a point estimate under F, e.g. the expectation μ_F.
- Logarithmic/Ignorance score (LOG/IGN): $S_{\text{LOG}}(F, y) = -\log f(y)$,
  where $f(\cdot)$ is the predictive density.
- Dawid-Sebastiani (DS): $S_{\text{DS}}(F, y) = \frac{(y-\mu_F)^2}{\sigma_F^2} + \log(\sigma_F^2)$.

Score expectations and proper scoring rules

- What functions of the predictive distributions are useful scores?
- We want to reward accurate (unbiased) and precise (small variance) predictions, but not at the expense of understating true uncertainty.
- First, we define the expectation of a score under a true distribution G as

$$S(F, G) = \mathsf{E}_{y \sim G}[S(F, y)]$$

Proper scores/scoring rules

A negatively oriented score is *proper* if it fulfils

$$S(F, G) \geq S(G, G).$$

A proper score that has equality of the expectations *only* when F and G are the same, $F(\cdot) \equiv G(\cdot)$, is said to be *strictly proper*.

The practical interpretation of this is that a proper score does not reward cheating; stating a lower (or higher) forecast/prediction uncertainty will not, on average, give a better score than stating the truth.

Absolute error and CRPS

Absolute error and Continuous Ranked Probability Score

- Absolute Error (AE): $S_{\mathsf{AE}}(F, y) = |y - \widehat{y}_F|$, where $\widehat{y}_F$ is a point estimate under F, e.g. the *median* $F^{-1}(1/2)$.
- CRPS: $S_{\mathsf{CRPS}}(F, y) = \int_{-\infty}^{\infty} \left[\mathbb{I}(y \leq x) - F(x)\right]^2 \, \mathrm{d}x$

Average scores

Average score

Given a collection of prediction/truth pairs, $\{(F_i, y_i), i = 1, \ldots, n\}$, define the *average* or *mean* score:

$$\overline{S}(\{(F_i, y_i), i = 1, \ldots, n\}) = \frac{1}{n} \sum_{i=1}^{n} S(F_i, y_i)$$

- When comparing prediction quality, we often look at the difference in average scores across the test data set.
- For modern, complex models with explicit spatial and temporal model components, the *pairwise* differences may be useful: For two prediction methods, F and F',

$$S_i^{\Delta}(F_i, F_i', y_i) = S(F_i, y_i) - S(F_i', y_i)$$

  We can have $\overline{S}^{\Delta} \approx 0$ at the same time as all $|S_i^{\Delta}| \gg 0$, if the two models/methods are both good, but e.g. at different spatial locations.
- How can we assess whether the score differences are indistinguishable?

How good are confidence/prediction interval procedures?

Tradeoffs for CIs

Desired properties for methods generating CIs for a quantity Y:

1. Appropriate *coverage* under the true distribution, G: $\mathsf{P}_G(Y \in CI_F) \geq 1 - \alpha$
2. Narrow intervals

- A wide prediction F helps with 1 but makes 2 difficult
- A narrow prediction F helps with 2 but makes 1 difficult

A proper score for interval predictions

The *Interval Score* For a CI (L_F, U_F) is defined by

$$S_{\mathsf{INT}}(F, y) = U_F - L_F + \frac{2}{\alpha}(L_F - y)\mathbb{I}(y < L_F) + \frac{2}{\alpha}(y - U_F)\mathbb{I}(y > U_F)$$

It is a proper scoring rule, consistent for equal-tail error probability intervals: $S(F, G)$ is minimised for the narrowest CI that has expected coverage $1 - \alpha$.

Proper scores

$$
\begin{aligned}
S_{\mathsf{SE}}(F,G) &= \mathsf{E}_{y\sim G}[S_{\mathsf{SE}}(F,y)] = \mathsf{E}_{y\sim G}[(y-\mu_F)^2] = \mathsf{E}_{y\sim G}[(y-\mu_G+\mu_G-\mu_F)^2] \\
&= \mathsf{E}_{y\sim G}[(y-\mu_G)^2 + 2(y-\mu_G)(\mu_G-\mu_F) + (\mu_G-\mu_F)^2] \\
&= \mathsf{E}_{y\sim G}[(y-\mu_G)^2] + 2(\mu_G-\mu_F)\mathsf{E}_{y\sim G}[y-\mu_G] + (\mu_G-\mu_F)^2 \\
&= \sigma_G^2 + (\mu_G-\mu_F)^2
\end{aligned}
$$

This is minimised when $\mu_F = \mu_G$. Therefore $S_{\mathsf{SE}}(F,G) \geq S_{\mathsf{SE}}(G,G) = \sigma_G^2$, so the score is proper. Is it strictly proper?

$$
\begin{aligned}
S_{\mathsf{DS}}(F,G) &= \mathsf{E}_{y\sim G}[S_{\mathsf{DS}}(F,y)] = \frac{\mathsf{E}_{y\sim G}[(y-\mu_F)^2]}{\sigma_F^2} + \log(\sigma_F^2) \\
&= \frac{\sigma_G^2 + (\mu_G-\mu_F)^2}{\sigma_F^2} + \log(\sigma_F^2)
\end{aligned}
$$

This is minimised when $\mu_F = \mu_G$ and $\sigma_F = \sigma_G$. Therefore $S_{\mathsf{DS}}(F,G) \geq S_{\mathsf{DS}}(G,G) = 1 + \log(\sigma_G^2)$, so the score is proper. Is it strictly proper?