Some Extensions to Bagging

Tom Heskes,
Foundation for Neural Networks,
Nijmegen, The Netherlands.

Two parts:

1. Weighted instead of equal averaging?
2. Other than sum-squared error measures?

Basics: sum-squared error, equal averaging

With target t, an ensemble of n estimators o_α, and error measure $(t - o_\alpha)^2$, we have

$$(t-m)^2 = \frac{1}{n}\sum_\alpha (t - o_\alpha)^2 - \frac{1}{n}\sum_\alpha (m - o_\alpha)^2$$

or

$$\text{error of average} = \text{average error} - \text{variance.}$$

The average m minimizes the variance ("ambiguity"):

$$m = \operatorname*{argmin}_{m'} \frac{1}{n}\sum_\alpha (o_\alpha - m')^2 = \frac{1}{n}\sum_\alpha o_\alpha \ .$$

Since variance ≥ 0, we always have

$$\text{error of average} \leq \text{average of error!!}$$

Weighted averaging (still sum-squared error)

Introducing weighting factors w_α, we still have

$$(t-m)^2 = \sum_\alpha w_\alpha (t-o_\alpha)^2 - \sum_\alpha w_\alpha (m-o_\alpha)^2$$

and thus

error of weighted average $\leq$ weighted average of error,

if we define

$$m = \operatorname*{argmin}_{m'} \sum_\alpha w_\alpha (m'-o_\alpha)^2 = \sum_\alpha w_\alpha o_\alpha \ ,$$

and (necessary!)

$$\sum_\alpha w_\alpha = 1.$$

Weighting factors better than $w_\alpha = 1/n$?

Maybe! Rewrite (Krogh and Vedelsby, NIPS 7)

$$(t-m)^2 = \sum_{\alpha} w_\alpha (t - o_\alpha)^2 - \frac{1}{2} \sum_{\alpha,\beta} w_\alpha w_\beta (o_\alpha - o_\beta)^2$$

and note that

- First term dependent on (unknown) target t.
- Second term independent of target.
- Quadratic form in terms of weighting factors w_α.

Optimizing the weighting factors

Taking averages:

$$E(t-m)^2 = \sum_\alpha w_\alpha E(t-o_\alpha)^2 - \frac{1}{2}\sum_{\alpha,\beta} w_\alpha w_\beta E(o_\alpha - o_\beta)^2.$$

For example, the average may be over a new set of test patterns with known inputs and unknown targets. Then we

- can simply compute the variance terms from the outputs of the models on the given inputs;
- but have to estimate the individual error terms, based on the performances on training data.

A quadratic programming problem

Now, suppose you can estimate $E(t - o_\alpha)^2$. Then finding optimal w_α is a straightforward quadratic programming problem with constraints

$$\sum_\alpha w_\alpha = 1$$

and, if your "prior belief" excludes negative weighting factors,

$$w_\alpha \geq 0 \quad \forall_\alpha \, .$$

Interpretation: search for models that are far apart and all have a small individual error. (For example, $e_1 = e_2 = e_3$; $d_{12} = 0$; $d_{13} = d_{23} \Rightarrow w_1 = w_2 = 1/4$; $w_3 = 1/2$.)

In practice, solutions tend to lie on the boundary of the hypercube (many $w_\alpha = 0$).

How do you estimate $E(t - o_\alpha)^2$?

No easy solution (as far as I know):

- *Separate test set, the same for all models.*
  Probably the least risky, but quite a waste of data.

- *Use the validation set (part of the data not in the bootstrap sample).*
  For free but tricky: some examples are more difficult to fit than others, i.e., strong negative correlation between training error and validation error.

It is possible to come up with more or less heuristic ways to correct for this effect.

Making it more robust

"Double bootstrapping":

1. Estimate not only $\chi_\alpha = E(t - o_\alpha)^2$ but also its uncertainty $\Delta\chi_\alpha$.
2. Bootstrap from the distribution of the estimated errors χ_α to obtain several estimates for the weighting factors w_α.
3. Average over these estimates.

Simulations show that this is indeed more robust, at the expense of having to keep more models.

End of part One

Is this all specific to sum-squared errors?

Not quite!

It is all more or less the same for any error measure that, using the appropriate probability model, can be derived from a Kullback-Leibler divergence.

Key trick: define the average over models in the right way.

Obvious relationship with bias/variance decompositions (NIPS 10:266; NC 10:1425; http://www.mbfys.kun.nl/~tom).

Little bit of notation

Variable to be estimated Y, target density $q(y)$ for the probability that $Y = y$, estimated densities $\hat{p}_\alpha(y)$, Kullback-Leibler divergence

$$K(q, \hat{p}) = \int dy\, q(y) \log \frac{q(y)}{\hat{p}(y)} \,.$$

Replace "density" by "distribution" and integrals by summations for discrete variables Y.

Example: densities are Gaussians with known variance and unknown means. Target mean t, estimated means o_α, brings us back to sum-squared error.

Generalization of sum-squared

$$(t-m)^2 = \frac{1}{n}\sum_\alpha (t-o_\alpha)^2 - \frac{1}{n}\sum_\alpha (m-o_\alpha)^2$$

with

$$m = \underset{m'}{\mathrm{argmin}} \sum_\alpha (m'-o_\alpha)^2$$

$$K(q,\bar{p}) = \frac{1}{n}\sum_\alpha K(q,\hat{p}_\alpha) - \frac{1}{n}\sum_\alpha K(\bar{p},\hat{p}_\alpha)$$

with

$$\bar{p} = \underset{p'}{\mathrm{argmin}} \sum_\alpha K(p',\hat{p}_\alpha)$$

The average model is the "logarithmic opinion pool"

$$\log \bar{p}(y) = \frac{1}{n}\sum_\alpha \log \hat{p}_\alpha(y) + \text{ constant }.$$

Logarithmic versus linear opinion pool

$$\bar{p}_{\log}(y) \propto \sqrt[n]{\prod_\alpha \hat{p}_\alpha(y)} \text{ (logarithmic) versus (linear) } \bar{p}_{\text{lin}}(y) = \frac{1}{n}\sum_\alpha \hat{p}_\alpha(y).$$

Both

- can be shown to obey $K(q,\bar{p}) \leq \frac{1}{n}\sum_\alpha K(q,\hat{p}_\alpha)$;
- yield the same decomposition for sum-squared error.

But only $\bar{p}_{\log}$

- minimizes the variance $\frac{1}{n}\sum_\alpha K(\bar{p},\hat{p}_\alpha)$;
- yields the decomposition for the loglikelihood-related Kullback.

$\bar{p}_{\text{lin}}$ minimizes the variance $\frac{1}{n}\sum_\alpha K(\hat{p}_\alpha,\bar{p})$, and yields a similar decomposition for the less-used (inverse) Kullback $K^{-1}(q,\hat{p}) = \int dy\, \hat{p}(y) \log \frac{q(y)}{\hat{p}(y)}$, with no link to loglikelihood.

An example: absolute error

Absolute error is proportional to loglikelihood if we take (fixed σ)

$$\hat{p}_\alpha(y) = \frac{1}{2\sigma} \exp\left[-\frac{|y - o_\alpha|}{\sigma}\right] .$$

The logarithmic opinion pool yields, with normalization Z,

$$\bar{p}(y) = \frac{1}{Z} \exp\left[-\frac{1}{n\sigma} \sum_\alpha |y - o_\alpha|\right] .$$

The decomposition of the loglikelihood becomes

$$-\sigma \log \bar{p}(t) = \frac{1}{n} \sum_\alpha |t - o_\alpha| + \sigma \log \int dy\, \exp\left[-\frac{1}{n\sigma} \sum_\alpha |y - o_\alpha|\right] .$$

Note that $-\sigma \log \bar{p}(t) \neq |t - m|$ for some m independent of t and that the resulting decomposition depends on σ!

Weighted averaging for Kullback divergence

Still we have

$$K(q,\bar{p}) = \sum_\alpha w_\alpha K(q,\hat{p}_\alpha) - \sum_\alpha w_\alpha K(\bar{p},\hat{p}_\alpha)\ ,$$

with

$$\bar{p}(y) = \underset{p'}{\operatorname{argmin}} \sum_\alpha w_\alpha K(p',\hat{p}_\alpha) \propto \exp\left[\sum_\alpha w_\alpha \log \hat{p}_\alpha(y)\right]\ .$$

Decomposes into $\begin{array}{c}\text{average error}\\ \text{variance}\end{array}$ which $\begin{array}{c}\text{does depend}\\ \text{does not depend}\end{array}$ on the target.

However, no quadratic form for w_α which makes (constraint) optimization a disaster.

Some rewriting

Substitute $q = \hat{p}_\beta$ and take $\sum_\beta w_\beta$ to find that

$$\sum_\beta w_\beta K(\hat{p}_\beta, \bar{p}) = \sum_{\alpha,\beta} w_\alpha w_\beta K(\hat{p}_\beta, \hat{p}_\alpha) - \sum_\alpha w_\alpha K(\bar{p}, \hat{p}_\alpha) \; ,$$

and thus

$$K(q, \bar{p}) = \sum_\alpha w_\alpha K(q, \hat{p}_\alpha) - \frac{1}{4} \sum_{\alpha,\beta} w_\alpha w_\beta [K(\hat{p}_\alpha, \hat{p}_\beta) + K(\hat{p}_\beta, \hat{p}_\alpha)] + \text{ rest} \, ,$$

where

$$\text{rest} \; = \frac{1}{2} \sum_\alpha w_\alpha [K(p_\alpha, \bar{p}) - K(\bar{p}, p_\alpha)] \; .$$

The variance is rewritten in a symmetric and an anti-symmetric part (the rest term).

Approximate quadratic form

Note that:

- For sum-squared error the rest term is zero (symmetric Kullback).
- The anti-symmetric rest term is an order of magnitude smaller than the symmetric term we take into account.

So we can derive weighting factors w_α by optimizing

$$EK(q,\bar{p}) \approx \sum_\alpha w_\alpha EK(q,\hat{p}_\alpha) - \frac{1}{4}\sum_{\alpha,\beta} w_\alpha w_\beta E[K(\hat{p}_\alpha,\hat{p}_\beta) + K(\hat{p}_\beta,\hat{p}_\alpha)]$$

under the appropriate constraints using quadratic programming.

Conclusions

Part One:

- The usual decomposition of sum-squared error allows for a straightforward algorithm for weighting different models.
- Be careful although for “overfitting” due to inaccurate estimates of the errors made by the individual models.

Part Two:

- The decomposition of sum-squared error is just a specific example of a more general decomposition of Kullback-Leibler divergences.
- The decomposition again leads to a quadratic form of the weighting factors if the asymmetry of the divergence is neglected.

An artefact?

Ask all these people to estimate something they don't know anything about, for example, the length of the Chinese emperor's nose.

Average over their estimates and you will, on average, have a better guess than by just asking one of them, no matter what the truth is!

Is this an artefact of our error definitions?