What is Deviance? (specifically in CART/rpart)

What is "Deviance," how is it calculated, and what are its uses in different fields in statistics?

In particular, I'm personally interested in its uses in CART (and its implementation in rpart in R).

I'm asking this since the wiki-article seems somewhat lacking and your insights will be most welcomed.

r | cart | rpart | deviance

asked Jan 26 '11 at 16:27

Tal Galili
6,265 11 66 124

3 Answers

Deviance and GLM

Formally, one can views deviance as a sort of distance between two probabilistic models; in an GLM context, it amounts to two times the log ratio of likelihoods between two nested models ℓ_1/ℓ_0 where ℓ_0 is the "smaller" model; that is, a linear restriction on model parameters (cf. the Neyman–Pearson lemma), as @suncoolsu said. As such, it can be used to perform *model comparison*. It can also be seen it as a generalization of the RSS used in OLS estimation (ANOVA, regression), for it provides a measure of *goodness-of-fit* of the model being evaluated when compared to the null model (intercept only). It works with LM too:

```
> x <- rnorm(100)
> y <- 0.8*x+rnorm(100)
> lm.res <- lm(y ~ x)
```

The residuals SS (RSS) is computed as $\hat{\varepsilon}^t\hat{\varepsilon}$, which is readily obtained as:

```
> t(residuals(lm.res))%*%residuals(lm.res)
         [,1]
[1,] 98.66754
```

or from the (unadjusted) R^2

```
> summary(lm.res)

Call:
lm(formula = y ~ x)

(...)

Residual standard error: 1.003 on 98 degrees of freedom
Multiple R-squared: 0.4234, Adjusted R-squared: 0.4175
F-statistic: 71.97 on 1 and 98 DF,  p-value: 2.334e-13
```

since $R^2 = 1 - \mathrm{RSS}/\mathrm{TSS}$ where TSS is the total variance. Note that it is directly available in an ANOVA table, like

```
> summary.aov(lm.res)
            Df Sum Sq Mean Sq F value    Pr(>F)
x            1 72.459  72.459  71.969 2.334e-13 ***
Residuals   98 98.668   1.007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Now, look at the deviance:

```
> deviance(lm.res)
[1] 98.66754
```

In fact, for linear models the deviance equals the RSS (you may recall that OLS and ML

estimates coincide in such a case).

Deviance and CART

We can see CART as a way to allocate already n labeled individuals into arbitrary classes (in a classification context). Trees can be viewed as providing a probability model for individuals class membership. So, at each node i, we have a probability distribution p_{ik} over the classes. What is important here is that the leaves of the tree give us a random sample n_{ik} from a multinomial distribution specified by p_{ik}. We can thus define the deviance of a tree, D, as the sum over all leaves of

$$D_i = -2 \sum_k n_{ik} \log(p_{ik}),$$

following Venables and Ripley's notations (MASS, Springer 2002, 4th ed.). If you have access to this essential reference for R users (IMHO), you can check by yourself how such an approach is used for splitting nodes and fitting a tree to observed data (p. 255 ff.); basically, the idea is to minimize, by pruning the tree, $D + \alpha \#(T)$ where $\#(T)$ is the number of nodes in the tree T. Here we recognize the *cost-complexity trade-off*. Here, D is equivalent to the concept of node impurity (i.e., the heterogeneity of the distribution at a given node) which are based on a measure of entropy or information gain, or the well-known Gini index, defined as $1 - \sum_k p_{ik}^2$ (the unknown proportions are estimated from node proportions).

With a regression tree, the idea is quite similar, and we can conceptualize the deviance as sum of squares defined for individuals j by

$$D_i = \sum_j (y_j - \mu_i)^2,$$

summed over all leaves. Here, the probability model that is considered within each leaf is a gaussian $\mathcal{N}(\mu_i, \sigma^2)$. Quoting Venables and Ripley (p. 256), "D is the usual scaled deviance for a gaussian GLM. However, the distribution at internal nodes of the tree is then a mixture of normal distributions, and so D_i is only appropriate at the leaves. The tree-construction process has to be seen as a *hierarchical refinement of probability models, very similar to forward variable selection in regression*." Section 9.2 provides further detailed information about `rpart` implementation, but you can already look at the `residuals()` function for `rpart` object, where "deviance residuals" are computed as the square root of minus twice the logarithm of the fitted model.

An introduction to recursive partitioning using the rpart routines, by Atkinson and Therneau, is also a good start. For more general review (including bagging), I would recommend

- Moissen, G.G. (2008). Classification and Regression Trees. *Ecological Informatics*, pp. 582-588.
- Sutton, C.D. (2005). Classification and Regression Trees, Bagging, and Boosting, in *Handbook of Statistics, Vol. 24*, pp. 303-329, Elsevier.

edited Nov 21 '11 at 20:31 — whuber ♦ — 117k · 12 · 209 · 419

answered Jan 27 '11 at 9:05 — chl ♦ — 34.1k · 6 · 103 · 222

+1 nice answer chl, which, as you say, complements mine so no problems there. Very nicely put. – Gavin Simpson Jan 27 '11 at 11:54

One minor thing - probably not best to use `lm.fit` in your example, as this is the name of the workhorse behind `lm`. – Gavin Simpson Jan 27 '11 at 14:57

@Gavin Oh, yes, that's right. Thx for reminding me that, I often make this mistake... – chl ♦ Jan 27 '11 at 15:03

+1 great answer! – suncoolsu Jan 27 '11 at 17:17

2 Just one remark: For linear regression Deviance is equal to the RSS because the normality assumption of errors implies RSS is same as LR test statistic which further implies Deviance is normally distributed, *irrespective of asymptotics*. This is just an expansion of (I think) terse comment of chl. – suncoolsu Jan 27 '11 at 18:11

It might be a bit clearer if we think about a perfect model with as many parameters as observations such that it explains all variance in the response. This is the saturated model. Deviance simply measures the difference in "fit" of a candidate model and that of the saturated model.

In a regression tree, the saturated model would be one that had as many terminal nodes (leaves) as observations so it would perfectly fit the response. The deviance of a simpler model can be computed as the node residual sums of squares, summed over all nodes. In other words, the sum of squared differences between predicted and observed values. This is the same sort of error (or deviance) used in least squares regression.

For a classification tree, residual sums of squares is not the most appropriate measure of lack of fit. Instead, there is an alternative measure of deviance, plus trees can be built minimising an entropy measure or the Gini index. The latter is the default in `rpart`. The Gini index is computed as:

$$D_i = 1 - \sum_{k=1}^{K} p_{ik}^2$$

where p_{ik} is the observed proportion of class k in node i. This measure is summed of all terminal i nodes in the tree to arrive at a deviance for the fitted tree model.

answered Jan 27 '11 at 8:47

Gavin Simpson
13.9k 2 38 73

(+1) Sorry, my post came later and I didn't notice yours. As I think they don't overlap too much, I will leave mine if you don't mind. – chl ♦ Jan 27 '11 at 10:02

So, *deviance* is a measure of goodness-of-fit, right? AFAIK, in regression, we have some statistics (such as RSS, R^2) to measure goodnees-of-fit; and in classification, we could use misclassification rate. Am I right? – loganecolss Mar 25 '14 at 12:12

Deviance is the likelihood-ratio statistic for testing the null hypothesis that the model holds agains the general alternative (i.e., the saturated model). For some Poisson and binomial GLMs, the number of observations N stays fixed as the individual counts increase in size. Then the deviance has a *chi-squared asymptotic null distribution*. The degrees of freedom = N - p, where p is the number of model parameters; i.e., it is equal to the numbers of free parameters in the saturated and unsaturated models. The deviance then provides a test for the model fit.

$$Deviance = -2[L(\hat{\mu}|\mathbf{y}) - L(\mathbf{y}|\mathbf{y})]$$

However, most of the times, you want to test if you need to drop some variables. Say there are two models M_1 and M_2 with p_1 and p_2 parameters, respectively, and you need to test which of these two is better. Assume M_1 is a special case of M_2 i.e. nested models.

In that case, the difference of deviance is taken:

$$\Delta Deviance = -2[L(\hat{\mu_1}|\mathbf{y}) - L(\hat{\mu_2}|\mathbf{y})]$$

Notice that the log likelihood of the saturated model cancels and the degree of freedom of $\Delta Deviance$ changes to $p_2 - p_1$. This is what we use most often when we need to test if some of the parameters are 0 or not. But when you fit `glm` in `R` the deviance output is for the saturated model vs the current model.

If you want to read in greater details: cf: Categorical Data Analysis by Alan Agresti, pp 118.

edited Jan 26 '11 at 22:20

mbq ♦
15.8k 7 44 93

answered Jan 26 '11 at 22:13

suncoolsu
4,468 19 36

@Tal, I don't use `rpart` and I will leave the answer to more experienced members of this forum. – suncoolsu Jan 26 '11 at 22:19

I think I've got the idea... But rpart prints deviance even for regression trees O.o – deps_stats Jan 26 '11 at 23:46

@deps_stats that deviance is node residual sums of squares summed over the terminal nodes of the tree. – Gavin Simpson Jan 27 '11 at 8:48