Statistics 101C Discussion Week 2

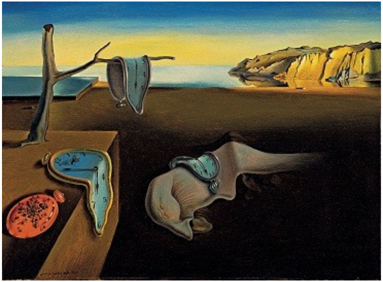
Brice Randolph

University of California, Los Angeles

brandolph@ucla.edu

April 5, 2016

Agenda



What will we be doing in discussion section today?

- Lecture Key Points
- ggplot2 examples with Hadley's code
- Review: R^2 , F test, MSE
- Features

The Dude Recommends

- Knitr
- LateX
- Python
- Familiarity with bash scripting (many tools are used from the command line)

The Dude Recommends

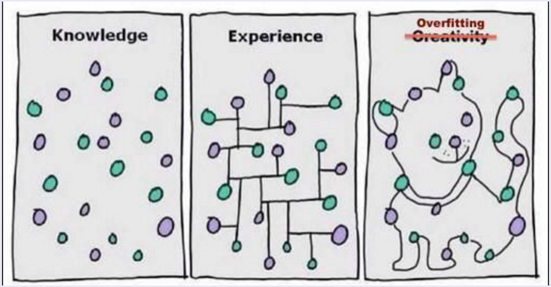
Install a good text editor:
I use vim for bash scripting and the IDE Xcode for C++/Swift(if you want to build an iphone app)

Bash commands you should know

- ls
- cd
- top
- mkdir
- ping
- ctrl-c (gets out of stuff) or ctrl-z
- q (try: $ps - A|less$) what does this do?
- clear
- ipython notebook

Modeling and Metrics

Signal and Noise



$$Y = f(x) + \epsilon$$

y = signal + noise

Overfitting

R^2 won't always help you

MSE = Reducible part + Irreducible part

$$= \mathbb{E}[(f(x) - \hat{f}(x))^2] + \sigma^2$$

At a point It can be decomposed further
In general how does a learning algorithm's flexibility relate to the bias and variance terms described above?

Risk function: $\mathcal{R}(Y_i, \hat{f}) = \mathbb{E}[(Y_i - \hat{f})^2 \mid X_i]$

$$\frac{\partial R(Y_i, \hat{f})}{\partial \hat{f}(X_i)} = \mathbb{E}[2(Y_i - \hat{f}) \mid X_i] = 0$$
$$\mathbb{E}[Y_i \mid X_i] = \mathbb{E}[\hat{f} \mid X_i]$$
$$\hat{f} = \mathbb{E}[Y_i \mid X_i]$$

$$
\begin{aligned}
\mathcal{R} &= \mathbb{E}[(Y_i - \hat{f}(X_i))^2] \\
&= \mathbb{E}[Y_i^2 - 2Y\hat{f}(X_i) + [\hat{f}(X_i)]^2] \\
&= \mathbb{E}[Y_i^2] - 2\,\mathbb{E}[Y_i]\hat{f}(X_i) + \mathbb{E}[[\hat{f}(X_i)]^2] \\
&= Var(Y_i) + \mathbb{E}[Y_i]^2 + Var(\hat{f}(X_i)) - 2\,\mathbb{E}[Y_i]\hat{f}(X_i) + \mathbb{E}[\hat{f}(X_i)]^2 \\
&= \sigma_{\epsilon_i}^2 + Var(\hat{f}(X_i)) + [\mathbb{E}[Y_i] - \mathbb{E}[\hat{f}(X_i)]]^2 \\
&= \sigma_{\epsilon_i}^2 + Var(\hat{f}(X_i)) + [f(X_i) - \mathbb{E}[\hat{f}(X_i)]^2 \\
&= \textbf{Irreducible Error + Variance of Estimator + Bias}^2 \textbf{ of Estimator}
\end{aligned}
$$

This decomposition shows that the risk of our estimator can be decomposed into three pieces:

- An error term coming from the underlying data generation process(that we can't change)
- A variance term that describes the variability in our estimate on new data
- A bias term that describes how well we fit the dataset used to train the model

Overfitting

Expected MSE measures the squared differences between observed and expected differences.

F statistic, MSE, R^2

Enormous Difference between training and testing error? What might this show?

Nested Models

What are they?
Two linear models are nested if one is obtained from the other by setting some parameters to zero or some other constraint on the parameters.

[(Restricted Model) Full Model]

We can compare nested models fit on the **same dataset** with the F test.

R^2

You can look these next slides over in detail if you need to.

Dense summary:
R^2 is the squared multiple correlation coefficient. It is also called the Coefficient of Determination. R^2 =RegSS/TotSS. It is the proportion of the variability in the response that is fitted by the model.
If a model has perfect predictability, R^2=1. If a model has no predictive capability, R^2=0. (In practice, R^2 is never observed to be exactly 0 the same way the difference between the means of two samples drawn from the same population is never exactly 0.) R, the multiple correlation coefficient and square root of R^2, is the correlation between the observed values (y), and the predicted values ($\hat{y}$).

More R^2

As additional variables are added to a regression equation, R^2 increases even when the new variables have no real predictive capability. The adjusted-R^2 is an R^2-like measure that avoids this difficulty. When variables are added to the equation, adj-R^2 doesn't increase unless the new variables have additional predictive capability.

Now, what does it mean?

F test

The F Value or F ratio is the test statistic used to decide whether the model as a whole has statistically significant predictive capability, that is, whether the regression SS is big enough, considering the number of variables needed to achieve it.

F is the ratio of the Model Mean Square to the Error Mean Square. Under the null hypothesis that the model has no predictive capability–that is, that all population regression coefficients are 0 simultaneously–the F statistic follows an F distribution with p numerator degrees of freedom and n-p-1 denominator degrees of freedom.

http://www.jerrydallal.com/lhsp/regout.htm Great Summary of regression diagnostics

Comparing F and R^2

11

If all the assumptions hold and you have the correct form for R^2 then the usual F statistic can be computed as $F = \frac{R^2}{1-R^2} \times \frac{\mathrm{df}_2}{\mathrm{df}_1}$. This value can then be compared to the appropriate F distribution to do an F test. This can be derived/confirmed with basic algebra.

share improve this answer

answered Apr 22 '13 at 17:44

Greg Snow

31.4k 40 99

Intuitively, I like to think that the result of the F-ratio first gives a yes-no response to the the question, 'can I reject H_0?' (this is determined if the ratio is much larger than 1, or the p-value < α).

Then if I determine I can reject H_0, R^2 then indicates the strength of the relationship between.

In other words, a large F-ratio indicates that there is a relationship. High R^2 then indicates how strong that relationship is.

share improve this answer

answered May 21 '13 at 9:16

Entropica

1 1

2 Mean Squared Errors

MSE of test data is unknown (future data - we don't have this right now)

MSE training - Example of picking a tailored suit

How did you fit the least squares line?

Test MSE vs Train MSE

What do they look like?

Which should be larger? Is this always the case?

Cross Validation

What is it used for?

Try to estimate how much worse your predictions will be on test data by comparing testing and training error performance.

Features?

The word **features** is all over the machine learning literature

Think of it as an abstraction of how we would think about facial features - how would you describe facial features?

Features

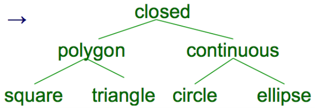
Standard feature types

- *nominal* (including Boolean)
  - no ordering among possible values
    e.g. *color* ϵ {*red*, *blue*, *green*} (vs. *color = 1000* Hertz)

- *linear* (or *ordinal*)
  - possible values of the feature are totally ordered
    e.g. *size* ϵ *{small, medium, large}* ← discrete
    weight ϵ [0…500] ← continuous

- *hierarchical*
  - possible values are partially *ordered* in an ISA hierarchy

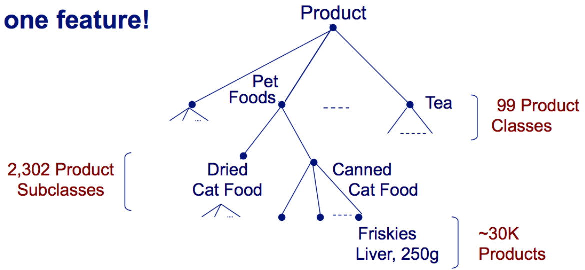
e.g. *shape* →



Feature hierarchy example

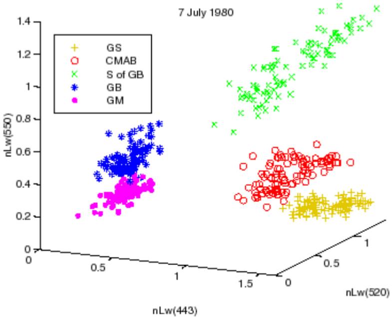
Lawrence et al., *Data Mining and Knowledge Discovery* 5(1-2), 2001

Structure of one feature!



Feature space

we can think of each instance as representing a point in a d-dimensional feature space where d is the number of features



example: optical properties of oceans in three spectral bands
[Traykovski and Sosik, *Ocean Optics XIV Conference Proceedings*, 1998]

Another view of the feature-vector representation: a single database table

	feature 1	feature 2	. . .	feature d	class
instance 1	0.0	small		red	true
instance 2	9.3	medium		red	false
instance 3	8.2	small		blue	false
. . .					
instance n	5.7	medium		green	true

The feature problem

Feature vector format is nice.

Unfortunately, real world data doesn't come in nice aligned feature vectors.

- Sequences: events in time, genomes, books
- Graphs: social networks, logistics (FedEx packages), sensor networks
- Relational databases: patient's health records are distributed over many tables.

ggplot2 examples- edited from Haldey Wickem

Open up R Studio:

Assume I don't know a thing about **shiny** (interactive web graphics in R).

Look at the following code and try to explain what the pieces do.

Feel free to tinker with the parameters:

http://shiny.rstudio.com/gallery/kmeans-example.html