problem_set_1

Steven Cao

1/15/2020

Part 1

Supervised Learning

The purpose of supervised learning is to train a model to make accurate predictions and accommodate incoming (i.e. new) observations. The "supervised" in "supervised learning" means teaching the model what makes for a good answer: that is possible because the answer is already known and can therefore be used to correct the model's mistakes. Every time mistakes are corrected, the model is also "nudged" in the right direction*. When this "nudging" process is iterated many times, error becomes minimal and the model becomes optimal.

This practise gives the model the ability to represent the "essence" of the data (e.g. what factors make an observation this class rather than that class), thereby allowing it to say something about unobserved events/entities with reasonable accuracy. In this approach, the data and its features are generally already well-understood and/or well-organised, and the goal is mainly to find out how these features are related to one another. Because these features are typically well-defined, the model can be specified to determine which features (independent variables) to relate to which other features (dependent/outcome variables, or class IDs). This can be used to classify observations into groups based on their features (classification), or infer (the value of) certain features of an observation based on its other features (regression).

The goal is to have a model that "captures the essence of the data" the best, and we measure this "essence-capturing" in terms of accuracy (or alternatively, error-minimisation). Assuming that the data was reliably sampled from the population (which is why the algorithm is only as good as the quality of data), a model which captures the essence of the data is one which also captures the essence of the population, and can thus be used to say something informative and accurate about the population.

*Although to qualify that claim, a complicated error function may have many local minima, and an error-minimising process like gradient descent may fall into a local minimum but not a global one, yielding a sub-optimal solution. I.e., it may not always be nudged in the absolutely *correct* direction.

Unsupervised Learning

The goals underlying unsupervised learning, on the other hand, are separate from supervised learning but can be complementary to them. It is usually pursued as a way of finding patterns and latent structure within the data in order to better approach or think about it. Reexpressing a dataset in terms of these patterns (e.g. relations between unclear features) can often simplify it. For example, dimension reduction techniques usually filter an overwhelming number of "raw features" into simpler-to-understand "higher-level features" that capture the essence of the "raw features" and their relations among each other. Such simplification is done often with the aim of visualisation, clearer interpretation, an aid to understanding, the exploration of data, and the generation of hypotheses (i.e. leads on the data).

One of the main differences between supervised and unsupervised learning is the lack of labeling on the data. Because we do not know the correct answers/values in advance, we cannot tell the model what is right or wrong, so we cannot "explicitly" nudge it in some "correct" direction: here, the concept of "correct/incorrect" does not exist. Thus, we cannot use error/accuracy to tune the model, and we cannot judge which model is "best" (for some given set of data) based on an accuracy measure. What we can do, on the other hand, is to let the model "decide for itself" (although constrained by our choice of algorithm) what would count as the best way to characterise data. In lieu of an error-based metric, we can use a distance-based metric to characterise the data: two observations that are closer together in their features are more alike, and are thus more likely to belong to the same class/cluster. (The best clusters generally minimise distance.)

In a nutshell, the point of unsupervised learning is to "give an understandable character to the data". The hope is that the model's characterisation of the data does relate to the data's essence in some way; thus the model is able to yield a way of seeing significance* in the data.

*I don't mean that in terms of statistical significance, but in the informal sense of seeing connections/structure.

Part 2, Problem a

```r
mtcars_lm1 <- lm(mpg ~ cyl, mtcars)
summary(mtcars_lm1)
```

```
##
## Call:
## lm(formula = mpg ~ cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9814 -2.1185  0.2217  1.0717  7.5186
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.8846     2.0738   18.27  < 2e-16 ***
## cyl          -2.8758     0.3224   -8.92 6.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.206 on 30 degrees of freedom
## Multiple R-squared:  0.7262, Adjusted R-squared:  0.7171
## F-statistic: 79.56 on 1 and 30 DF,  p-value: 6.113e-10
```

The coefficient for the `cyl` variable is -2.8758, with a t-statistic of -8.92. That means that for every 1 "unit" increase in the number of engine cylinders, the vehicle will generally get 2.876 less miles to the gallon. Additionally, this relationship is significant (because of the high t-statistic), so it is unlikely to be due to pure chance.

The constant coefficient is 37.8846, with a t-statistic of 18.27. Although the constant coefficient is usually not very interpretable, in this case, it *nominally* means that a vehicle which somehow has no cylinders would be predicted to yield 37.8846 miles to the gallon.

Part 2, Problem b

$$mpg_i = \beta_0 + \beta_1 cyl_i + \varepsilon_i$$

Part 2, Problem c

```r
mtcars_lm2 <- lm(mpg ~ cyl + wt, mtcars)
summary(mtcars_lm2)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2893 -1.5512 -0.4684  1.5743  6.1004
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
## cyl          -1.5078     0.4147  -3.636 0.001064 **
## wt           -3.1910     0.7569  -4.216 0.000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.568 on 29 degrees of freedom
## Multiple R-squared:  0.8302, Adjusted R-squared:  0.8185
## F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12
```

The t-statistics of both the `cyl` and `wt` variables both indicate significance (i.e. that the relationship is likely not due to pure chance).

The coefficient for the `cyl` variable is now -1.5078, which is less than it was before. This means that for every additional cylinder the vehicle has, the vehicle will generally get 1.508 less miles to the gallon. The coefficient for the `wt` variable is -3.1910, which is greater than the coefficient for `cyl`. This means that for every additional kilogram the vehicle weighs, it will generally get 3.191 less miles to the gallon: thus, the effect of a unit difference in weight is larger than the effect of a unit difference in cylinder.

Part 2, Problem d

```r
mtcars_lm3 <- lm(mpg ~ cyl + wt + cyl*wt, mtcars)
summary(mtcars_lm3)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + cyl * wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2288 -1.3495 -0.5042  1.4647  5.2344
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.3068     6.1275   8.863 1.29e-09 ***
## cyl          -3.8032     1.0050  -3.784 0.000747 ***
## wt           -8.6556     2.3201  -3.731 0.000861 ***
## cyl:wt        0.8084     0.3273   2.470 0.019882 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.368 on 28 degrees of freedom
## Multiple R-squared:  0.8606, Adjusted R-squared:  0.8457
## F-statistic: 57.62 on 3 and 28 DF,  p-value: 4.231e-12
```

All three independent terms (`cyl`, `wt`, and `cyl:wt`) have t-statistics which indicate significance. The coefficients, respectively, are -3.8032, -8.6556, and 0.8084. Both `cyl` and `wt` have larger coefficients than in the previous model (which did not include the interaction term). The interaction term, on the other hand, has a positive coefficient. Overall, this indicates that while increments in either feature alone reduce the vehicle's mpg, there is an interaction effect which can receive a substantial interpretation: namely, that vehicles with more cylinders will also tend to weigh more (which suggests that `cyl` and `wt` had explained some of the same variance in `mpg`). The theoretical assertion of including such an interactive term is that both independent variables should covary in some meaningful way (i.e. that they are related to one another).

Part 3, Problem a

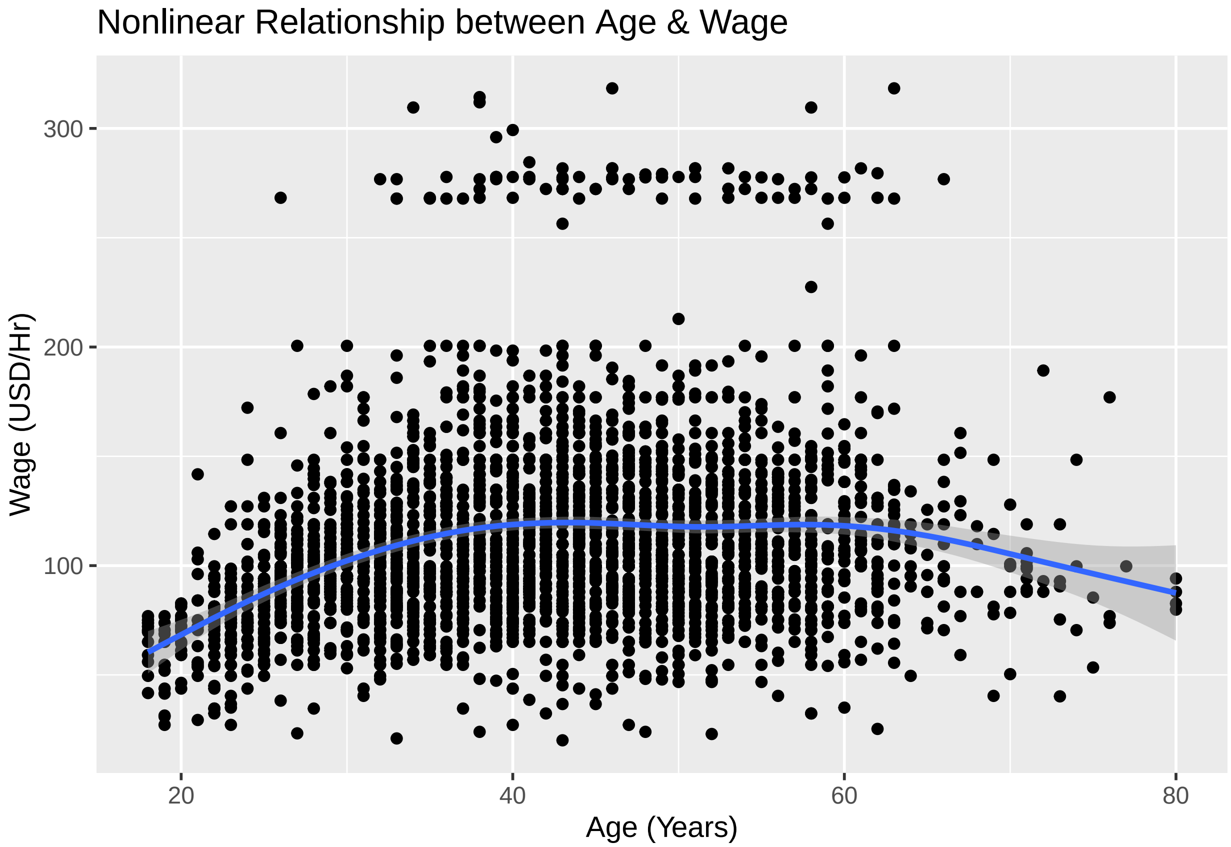
```r
wd_model <- lm(wage ~ age + I(age^2), wage_data)
summary(wd_model)
```

```
##
## Call:
## lm(formula = wage ~ age + I(age^2), data = wage_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.126 -24.309  -5.017  15.494 205.621
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.425224   8.189780  -1.273    0.203
## age           5.294030   0.388689  13.620   <2e-16 ***
## I(age^2)     -0.053005   0.004432 -11.960   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.99 on 2997 degrees of freedom
## Multiple R-squared:  0.08209,    Adjusted R-squared:  0.08147
## F-statistic:   134 on 2 and 2997 DF,  p-value: < 2.2e-16
```

The coefficients for the intercept, age, and age-squared variables were -10.425, 5.294, and -0.053 respectively. The t-statistics for both independent variables (but not the constant) indicated significance. While coefficients for nonlinear models are typically not straightforward to interpret (e.g. that an increase in 1 year of age corresponds to some intuitively-understandable change in wage), we can generally say that the shape of the function resembles an upside-down parabola. This means that wage will increase at first (with increasing age), before leveling out and eventually decreasing.

Part 3, Problem b

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Part 3, Problem c

People who are middle-aged - usually old enough to be in a senior-level position (and as particularly pronounced by the sparse cluster of high-earners at the top of the graph) - make more money overall than people who are younger or older. Such persons are likely to be working at a lower wage in an entry-level position (when they are younger) or are closer to retirement (and having less income because they are not working). By fitting a polynomial regression, we are asserting that there is not a linear relationship between two variables - in this case, that there is not a straightforward and monotonic relationship between age and wage.

Part 3, Problem d

Polynomial regression runs the danger of overfitting while overall being more difficult to meaningfully interpret (when it comes to reading out the meaning of the model's coefficients). Linear regression, on the other hand, does not run the danger of overfitting and is generally straightforward to interpret (i.e. that one unit of increase in some independent variable yields some corresponding increase in the dependent variable). Additionally, nonlinear regression is typically fitted as a result of iteration ("trial-and-error") while linear regression has a closed-form solution.