MLPS1

Ying Zhou

2020/1/16/

Q1

$X's$ are predictor feature measurements (inputs, independent variables, predictors, features) and Y is some observed quantitative response (outputs, dependent variable).

In supervised machine learning, we use the labeled data (both inputs and outputs are used). Each observation of X_i has an associated response measurement Y_i. We assume that the generalized relationship between Y and X is $Y = f(X) + \epsilon$ where $f(.)$ is some fixed unknown function of X and ϵ is a random error term that is independent of X with mean 0. We aim to learn and estimate the function that best approximates the relationship between inputs and outputs of given data, and aim to predict the outcome of future instance based on the estimated function. There are two kinds of prediction tasks: regression and classification. When we predict quantitative outputs, we do regression. When we predict qualitative outputs, we do classification. In the data generating process, first we need to create a training dataset from raw data and labels. The training dataset will contain the outputs and corresponding features that determine the outputs. During learning algorithm (e.g. linear regression, random forest, classification trees), the machine selects the optimal model among many models. Then if we give the machine new data/new inputs, it can generate the label. Supervised learning solves many kinds of real-world computation problems.

In unsupervised machine learning, we use the unlabeled data. That is, only input data is needed. For each observation of X_i, there is no associated response measurement Y_i. We aim to describe the associations and patterns among a set of input measures, i.e. to learn the inherent structure or pattern in a collection of uncategorized data. The common tasks in unsupervised learning are clustering, representation learning, and density estimation. Algorithms include k-means clustering, principal component analysis, etc. Unsupervised learning is useful for exploratory analysis because it identifies the structure of data. It is also useful for dimensionality reduction which is used to eliminate the insignificant independent variables. Unsupervised learning identifies the features for categorization.

It is easier to get unlabeled data from a computer than labeled data. Supervised learning is a simpler method, while unsupervised learning is computationally complex. Supervised learning has a clear measurement of success that compares the performance/effectiveness of different models, but there is no specific way in most unsupervised learning methods.

Linear Regression

(a)

```
model1 <- lm(mpg ~ cyl, data = mtcars)
summary(model1)
```

```
##
## 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
```

(b)

mpg=37.885-2.876cyl

(c)

```
model2 <- lm(mpg ~ cyl+wt, data = mtcars)
summary(model2)
```

```
##
## 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 coefficient of cyl is statistically significant in both (a) and (c). The coefficient size of cyl is smaller in (c) than in (a), which means that the negative impact of cyl on mpg is not that large as (a) states because weight also has negative impact on mpg. The model in (c) is more effective and more predictive than that in (a), since adjusted R-squared in (c) is greater than (a).

(d)

```
wtcyl<- mtcars$wt*mtcars$cyl
model3 <- lm(mpg ~ cyl+wt+wtcyl, data = mtcars)
summary(model3)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + wtcyl, 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 ***
## wtcyl         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 coefficients are different. The coefficent size of cyl and wt are larger in (d) than in (c). The direction of impact of cyl and that of wt are the same, both are negative in (c) and (d). cyl is statistically significant in (a)(c)(d) under 95% confidence level, and wt is statistically significant in (c)(d) under 95% confidence level. By including a multiplicative interaction term in the function, we assert that the effect of cylinders on mpg depends on vehicle weight.

Nonlinear Regression

(a)

```
wagedata <- read.csv("C:/Users/zhouy/Desktop/Uchicago/uchicourse/Intro to Machine Learning/PS/PS1/wage_
library("ggplot2")
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

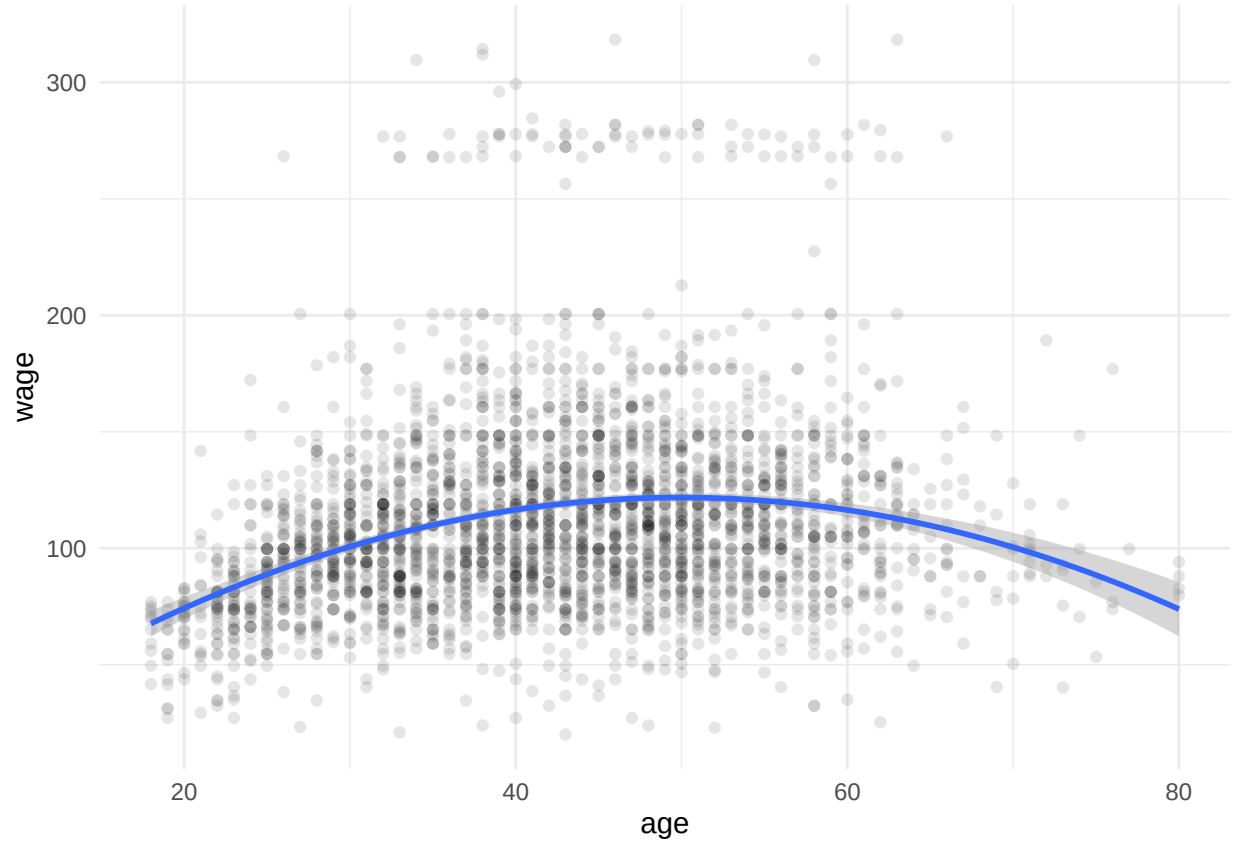
```
fit <- lm(wage~poly(age,2,raw=TRUE),data=wagedata)
summary(fit)
```

```
##
## Call:
## lm(formula = wage ~ poly(age, 2, raw = TRUE), data = wagedata)
##
## 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
## poly(age, 2, raw = TRUE)1    5.294030   0.388689  13.620   <2e-16 ***
## poly(age, 2, raw = TRUE)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
```

(b)

```
ggplot(data=wagedata,aes(age, wage)) +
  geom_point(alpha = 0.1)+
  geom_smooth(method=lm, formula=y~poly(x,2),level=0.95)+
  theme_minimal()
```



(c)

Before around 50 years old, wage is increasing with age. After 50 years old, wage is decreasing with age. We assert that the relationship between wage and age is non-linear.

(d)

Polynomial regression gives a nonlinear model, but linear regression gives a linear model. Polynomial regression has the greater number of regressors for one feature, and so more susceptible to overfitting. Linear regression only gives monotonicity (nomotonically increasing or monotonically decreasing), but nonlinear regression captures more precise changes of outputs with inputs.