caret 팩키지

컨닝쪽지(Cheat Sheet)

모형설정 (Specifying the Model)

모형에 변수를 설정하는 몇가지 가능한 구문:

```
train(y ~ x1 + x2, data = dat, ...)
train(x = predictor_df, y = outcome_vector, ...)
train(recipe_object, data = dat, ...)
```

- 단, rfe, sbf, gafs, safs 는 x/y 인터페이스만 지원.
- train 모형공식 함수는 formula method will **항상** 가변수(dummy variable)를 생성시킨다.
- train 함수의 x/y 인터페이스는 가변수를 생성시키지 않는다. (하지만, 기저 모형함수는 생성시킬 수도 있다).

다음 사항은 항상 기억한다:

- 데이터프레임에 칼럼명은 필수.
- 분류기 결과는 자료구조로 요인(factors)형 (0/1 혹은 정수는 안됨).
- 범주 수준에 적법한 R 명칭 사용 (“0”/”1”은 안됨)
- 모형학습시 동일한 재표집 표본을 얻으려면 train 함수호출하기 전에 난수 초기값(seed)를 설정.
- 결측값을 채워넣으려면, train 함수 선택옵션에 na.action = na.pass 명기. 마찬가지로 동일한 선택옵션을 결측값이 포함된 예측할 데이터에도 적용하여 예측.

train 함수 기저모형에 모형인자를 전달하려면 내부에 *생략(ellipses)*해서 전달할 수 있다.

```
train(y ~ ., data = dat, method = "rf",
    # `randomForest` 선택옵션:
    importance = TRUE)
```

병렬처리 (Parallel Processing)

병렬처리하는데 foreach 팩키지를 사용할 수 있다. train 코드는 바뀌지 않지만, “do” 팩키지는 먼저 호출되어야 한다.

```
# 맥 혹은 리눅스
library(doMC)
registerDoMC(cores=4)
```

```
# 윈도우
library(doParallel)
cl <- makeCluster(2)
registerDoParallel(cl)
```

parallel::detectCores 함수는 코어숫자 파악에 도움이 됨.

전처리 (Preprcoessing)

preProc 옵션을 변수변환, 필터링, 기타 다른 연산작업을 *예측변수(predictors)*에 적용시킬 수 있다.

```
train(, preProc = c("method1", "method2"), ...)
```

전처리 함수로 다음이 포함된다:

- 예측변수 정규화: "center", "scale", "range".
- 예측변수 변수변환: "BoxCox", "YeoJohnson", "expoTrans".
- 결측값 대체: "knnImpute", "bagImpute", "medianImpute".
- 필터링: "corr", "nzv", "zv", "conditionalX".
- 그룹 변환: "pca", "ica", "spatialSign".

train 함수가 연산우선순위를 결정한다; 전처리 함수가 선언되는 순서는 문제가 되지 않음.

recipes 팩키지에는 전처리 작업에 대한 훨씬 다양한 기법이 담겨있음.

선택옵션추가 (Adding Options)

대다수 train 선택옵션은 trainControl 함수를 별도 사용해서 지정할 수 있다.

```
train(y ~ ., data = dat, method = "cubist",
    trControl = trainControl(<선택옵션>))
```

재표집 선택옵션 (Resampling Options)

재표집 방법을 지정하는데 trainControl 함수를 사용한다.

```
trainControl(method = <method>, <options>)
```

함수와 선택옵션은 다음과 같다:

- "cv" : K-배 교차검증 (number: 배수를 지정).
- "repeatedcv": 반복 교차검증 (repeats: 반복횟수).
- "boot": 붓스트랩 (number: 되풀이 횟수).
- "LGOCV": 그룹 관측점 제거 방법 (number, p 는 선택옵션).
- "LOO": 단일 관측점 제거 교차검증.
- "oob" out-of-bag 재표집 (일부 모형 한정).
- "timeslice": 시계열 데이터 (선택옵션: initialWindow, horizon, fixedWindow, skip).

성능측도 (Perfomance Metric)

모형요약 방법을 선택하는데, trainControl 함수가 다시 사용된다.

```
trainControl(summaryFunction = <R function>,
        classProbs = <logical>)
```

자체제작 R 함수를 사용할 수 있지만, caret 팩키지에 일부가 내장되어 있음: defaultSummary (모형 정확도, RMSE, 등), twoClassSummary (ROC 곡선), prSummary (정보 검색). 마지막 두함수를 사용하려면, classProbs 를 TRUE로 설정해야 함.

격자탐색 (Grid Search)

train 함수가 튜닝 모수값을 결정하는데, 모수를 튜닝하는데 얼마나 많은 값을 탐색할지 tuneLength 선택옵션으로 제어한다.

또 다른 방법으로, 튜닝 모수값으로 tuneGrid 인자를 활용하는 것도 가능하다:

```
grid <- expand.grid(alpha = c(0.1, 0.5, 0.9),
                    lambda = c(0.001, 0.01))
train(x = x, y = y, method = "glmnet",
    preProc = c("center", "scale"),
    tuneGrid = grid)
```

임의 탐색 (Random Search)

모수 튜닝할 때, train 함수는 탐색할 범위에 포함된 튜닝모수값을 임의로 생성하는 기능이 있다. tuneLength 를 통해 탐색할 전체 조합수를 제어한다. 임의탐색은 다음과 같이 사용한다:

```
trainControl(search = "random")
```

하위 표집 (Subsampling)

클래스 불균형이 심한 경우, 모형적합 전에 클래스 균형을 맞추는데 train 함수 하위표집 기능을 사용한다:

```
trainControl(sampling = "down")
```

다른 값으로 "up", "smote", "rose" 지정이 가능하지만, "smote", "rose" 의 경우 추가로 팩키지 설치가 필요하다.