SHAO_mgcv_gamboost_gbm

shao

Friday, May 15, 2015

Prepare the data

```
library(Metrics)
library(data.table)   ## load data in quickly with fread
x <- fread("E:/Dropbox/kaggle/West Nile Virus Prediction/data/train.csv")
test <- fread("E:/Dropbox/kaggle/West Nile Virus Prediction/data/test_GAM.csv")

## prep the species column by moving the test-only
## UNSPECIFIED CULEX to CULEX ERRATICUS, and re-doing the levels
## logistic regression will complain otherwise
vSpecies<-c(as.character(x$Species),as.character(test$Species))
vSpecies[vSpecies=="UNSPECIFIED CULEX"]<-"CULEX ERRATICUS"
vSpecies[-which(vSpecies == "CULEX PIPIENS" |
                  vSpecies == "CULEX PIPIENS/RESTUANS" |
                  vSpecies == "CULEX RESTUANS")] = "CULEX OTHER"
vSpecies<-factor(vSpecies,levels=unique(vSpecies))

## data.table syntax for adding a column; could overwrite the existing column as well
x[,Species2:=factor(vSpecies[1:nrow(x)],levels=unique(vSpecies))]
test[,Species2:=factor(vSpecies[(nrow(x)+1):length(vSpecies)],levels=unique(vSpecies))]

## also add some fields for components of the date using simple substrings
x[,dMonth:=as.numeric(paste(substr(x$Date,6,7)))]
x[,dYear:=as.numeric(paste(substr(x$Date,1,4)))]
x$Date = as.Date(x$Date, format="%Y-%m-%d")
xsDate = as.Date(paste0(x$dYear, "0101"), format="%Y%m%d")
x$dWeek = as.numeric(paste(floor((x$Date - xsDate + 1)/7)))

test[,dMonth:=as.numeric(paste(substr(test$Date,6,7)))]
test[,dYear:=as.numeric(paste(substr(test$Date,1,4)))]
test$Date = as.Date(test$Date, format="%Y-%m-%d")
tsDate = as.Date(paste0(test$dYear, "0101"), format="%Y%m%d")
test$dWeek = as.numeric(paste(floor((test$Date - tsDate + 1)/7)))
## train set
x$TrapNumber <- as.integer(substr(as.character(x$Trap), 2, 4))

## test set
test$TrapNumber <- as.integer(substr(as.character(test$Trap), 2, 4))


# we'll set aside 2011 data as test, and train on the remaining
my.x = data.frame(x[,list(WnvPresent, dYear,dMonth,dWeek, Species2, Latitude, Longitude
                          ,Block, TrapNumber,AddressAccuracy,NumMosquitos)])
x1<-my.x[x$dYear!=2011,]
xcv<-my.x[x$dYear==2011,]
```

GAM (Generalized Addtive Model)

```
require(mgcv)
```

```
## Loading required package: mgcv
## Loading required package: nlme
## This is mgcv 1.8-6. For overview type 'help("mgcv-package")'.
```

```
fitCv1 = gam(WnvPresent ~ s(TrapNumber)+ s(dWeek) +Species2+s(Block)+
               s(NumMosquitos)+s(Latitude, Longitude)+s(dWeek,Species2,bs="fs")
             , data = x1,  family = binomial)
##s(dWeek,Species2,bs="fs")+s(NumMosquitos,Species2,bs="fs")+ s(Latitude, Longitude)+
##s(TrapNumber)+s(dWeek) +Species2
p1<-predict(fitCv1, newdata = xcv, type = "response")
summary(fitCv1)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## WnvPresent ~ s(TrapNumber) + s(dWeek) + Species2 + s(Block) +
##     s(NumMosquitos) + s(Latitude, Longitude) + s(dWeek, Species2,
##     bs = "fs")
##
## Parametric coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -4.307e+00  1.845e-01 -23.347   <2e-16 ***
## Species2CULEX RESTUANS  3.246e-02  2.052e-01   0.158   0.8743
## Species2CULEX PIPIENS   2.936e-01  1.334e-01   2.200   0.0278 *
## Species2CULEX OTHER    -4.119e+01  4.733e+06   0.000   1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                              edf Ref.df  Chi.sq  p-value
## s(TrapNumber)          1.003469  1.006   0.627   0.4303
## s(dWeek)               4.136517  5.058 204.930  < 2e-16 ***
## s(Block)               2.595161  3.163   8.464   0.0428 *
## s(NumMosquitos)        4.220519  5.151 272.077  < 2e-16 ***
## s(Latitude,Longitude) 15.380602 20.007  61.863 3.67e-06 ***
## s(dWeek,Species2)      0.000512 26.000   0.000   1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.164   Deviance explained = 26.3%
## UBRE = -0.66452  Scale est. = 1         n = 8452
```

```
## check for a reasonable AUC of the model against unseen data (2011)
auc(xcv$WnvPresent,p1)
```

```
## [1] 0.8921233
```

```
fitCv2 = gam(WnvPresent ~  s(Block)+s(NumMosquitos)+s(Latitude, Longitude)
             +te(dWeek,Species2,bs="fs"), data = x1,  family = binomial)
p2<-predict(fitCv2, newdata = xcv, type = "response")
summary(fitCv2)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## WnvPresent ~ s(Block) + s(NumMosquitos) + s(Latitude, Longitude) +
##     te(dWeek, Species2, bs = "fs")
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.063      2.336  -2.595  0.00946 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                           edf Ref.df Chi.sq p-value
## s(Block)                2.243  2.765  12.93 0.00414 **
## s(NumMosquitos)         4.289  5.231 262.55 < 2e-16 ***
## s(Latitude,Longitude)   3.695  4.894  62.68 4.1e-12 ***
## te(dWeek,Species2)     12.038 19.000 248.19 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.158   Deviance explained = 25.7%
## UBRE = -0.66347  Scale est. = 1         n = 8452
```

```
auc(xcv$WnvPresent,p2)
```

```
## [1] 0.8972362
```

```
fitCv3 = gam(WnvPresent ~  s(Block)+s(NumMosquitos)+te(Latitude, Longitude)
             +te(dWeek,Species2,bs="fs"), data = x1,  family = binomial)
p3<-predict(fitCv3, newdata = xcv, type = "response")
summary(fitCv3)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## WnvPresent ~ s(Block) + s(NumMosquitos) + te(Latitude, Longitude) +
##     te(dWeek, Species2, bs = "fs")
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)   -6.059      2.284  -2.653  0.00798 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                            edf Ref.df  Chi.sq p-value
## s(Block)                 2.243  2.776   7.533   0.048 *
## s(NumMosquitos)          4.267  5.206 259.828 < 2e-16 ***
## te(Latitude,Longitude) 11.457 13.491  72.503 5.1e-10 ***
## te(dWeek,Species2)     12.005 19.000 249.031 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.162   Deviance explained = 26.1%
## UBRE = -0.66357  Scale est. = 1         n = 8452
```

```
auc(xcv$WnvPresent,p3)
```

```
## [1] 0.8951366
```

```
fitCv4 = gam(WnvPresent ~  s(Block)+s(Latitude, Longitude)+te(dWeek,Species2,bs="fs")
             +te(NumMosquitos,Species2,bs="fs"), data = x1,  family = binomial)
p4<-predict(fitCv4, newdata = xcv, type = "response")
summary(fitCv4)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## WnvPresent ~ s(Block) + s(Latitude, Longitude) + te(dWeek, Species2,
##     bs = "fs") + te(NumMosquitos, Species2, bs = "fs")
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.208      5.086  -0.827    0.408
##
## Approximate significance of smooth terms:
##                               edf Ref.df  Chi.sq  p-value
## s(Block)                    2.651  3.233   8.156   0.0516 .
## s(Latitude,Longitude)      14.849 19.345  67.661 2.98e-07 ***
## te(dWeek,Species2)         11.155 19.000 211.492  < 2e-16 ***
## te(NumMosquitos,Species2)   9.486 16.000 264.990  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.166   Deviance explained = 26.6%
## UBRE = -0.66391  Scale est. = 1         n = 8452
```

```
auc(xcv$WnvPresent,p4)
```

```
## [1] 0.8885038
```

```
fitCv5 = gam(WnvPresent ~  Species2+s(Block,k=3,bs="re")+s(NumMosquitos,k=3)
             +s(Latitude, Longitude,k=4,bs="sos")
             +te(dWeek,Species2,bs="fs",k=3), data = x1,  family = binomial,gamma=2)
p5<-predict(fitCv5, newdata = xcv, type = "response")
summary(fitCv5)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## WnvPresent ~ Species2 + s(Block, k = 3, bs = "re") + s(NumMosquitos,
##     k = 3) + s(Latitude, Longitude, k = 4, bs = "sos") + te(dWeek,
##     Species2, bs = "fs", k = 3)
##
## Parametric coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -7.184e+00  7.293e+00  -0.985    0.325
## Species2CULEX RESTUANS  8.742e-01  1.033e+01   0.085    0.933
## Species2CULEX PIPIENS  -2.282e+00  1.034e+01  -0.221    0.825
## Species2CULEX OTHER    -3.565e+01  4.733e+06   0.000    1.000
##
## Approximate significance of smooth terms:
##                           edf Ref.df  Chi.sq p-value
## s(Block)               0.6039  1.000   3.478  0.0163 *
## s(NumMosquitos)        1.9642  1.998 271.671  <2e-16 ***
## s(Latitude,Longitude) 2.0136  3.000  66.743  <2e-16 ***
## te(dWeek,Species2)     5.8825  7.000 163.468  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.15   Deviance explained = 24.6%
## UBRE = -0.65739  Scale est. = 1         n = 8452
```

```
auc(xcv$WnvPresent,p5)
```

```
## [1] 0.9067505
```

```
anova(fitCv1,fitCv2,fitCv3,fitCv4,fitCv5,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: WnvPresent ~ s(TrapNumber) + s(dWeek) + Species2 + s(Block) +
##     s(NumMosquitos) + s(Latitude, Longitude) + s(dWeek, Species2,
##     bs = "fs")
## Model 2: WnvPresent ~ s(Block) + s(NumMosquitos) + s(Latitude, Longitude) +
##     te(dWeek, Species2, bs = "fs")
## Model 3: WnvPresent ~ s(Block) + s(NumMosquitos) + te(Latitude, Longitude) +
##     te(dWeek, Species2, bs = "fs")
## Model 4: WnvPresent ~ s(Block) + s(Latitude, Longitude) + te(dWeek, Species2,
##     bs = "fs") + te(NumMosquitos, Species2, bs = "fs")
```

```
## Model 5: WnvPresent ~ Species2 + s(Block, k = 3, bs = "re") + s(NumMosquitos,
##     k = 3) + s(Latitude, Longitude, k = 4, bs = "sos") + te(dWeek,
##     Species2, bs = "fs", k = 3)
##   Resid. Df Resid. Dev       Df Deviance  Pr(>Chi)
## 1    8420.7     2772.8
## 2    8428.7     2797.8  -8.0713  -25.006  0.001628 **
## 3    8421.0     2781.6   7.7064   16.284  0.033187 *
## 4    8412.9     2762.3   8.1691   19.216  0.015109 *
## 5    8437.5     2837.9 -24.6768  -75.587 4.552e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GAMboost (Generalized Addtive Model Boosting)

```
## GAMboost modelling
require(mboost)
```

```
## Loading required package: mboost
## Loading required package: parallel
## Loading required package: stabs
## This is mboost 2.4-2. See 'package?mboost' and 'news(package  = "mboost")'
## for a complete list of changes.
```

```
fitCv = gamboost(as.factor(WnvPresent) ~ dWeek +bols(Species2)+
                   bbs(dWeek)%X%bols(Species2)+Latitude+Longitude+Block+
                   NumMosquitos, data = x1,  control = boost_control(mstop = 200),
                 family = Binomial())

p2<-predict(fitCv, newdata = xcv, type = "response")
auc(xcv$WnvPresent,p2)
```

```
## [1] 0.8961556
```

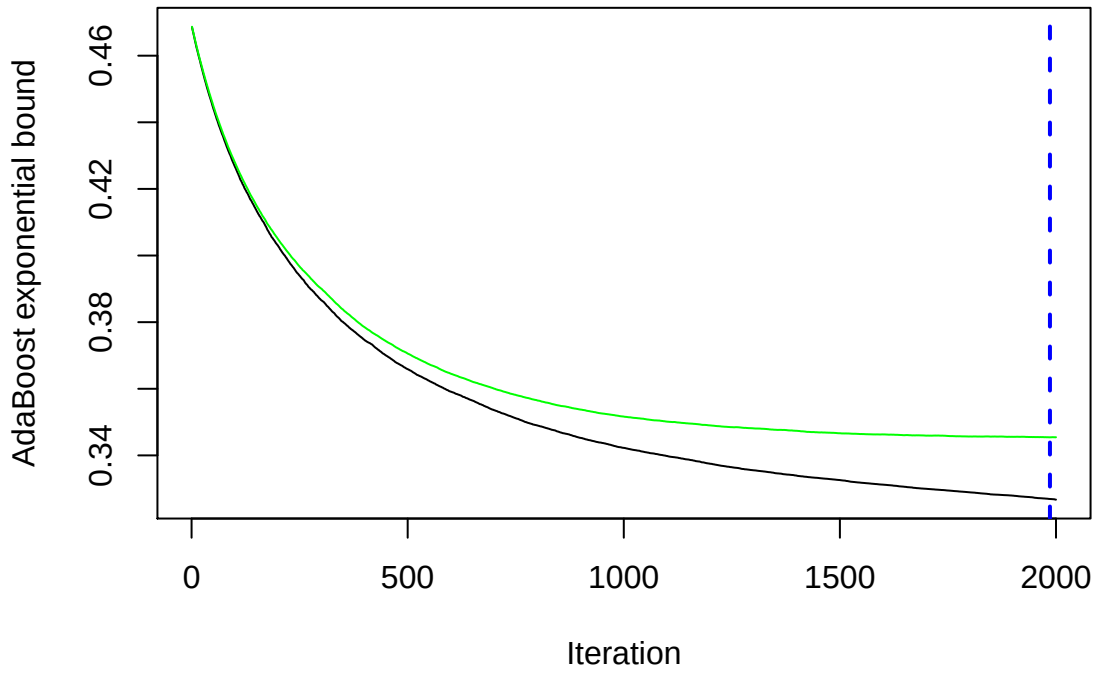
Gradient Boosting

```
require(gbm)
```

```
## Loading required package: gbm
## Loading required package: survival
## Loading required package: lattice
## Loading required package: splines
## Loaded gbm 2.1.1
```

```
set.seed(2)
fitCv = gbm(WnvPresent ~ dYear+dWeek +Species2+Latitude+Longitude+TrapNumber+
              NumMosquitos, data = x1, n.trees = 2000, interaction.depth = 2,
```

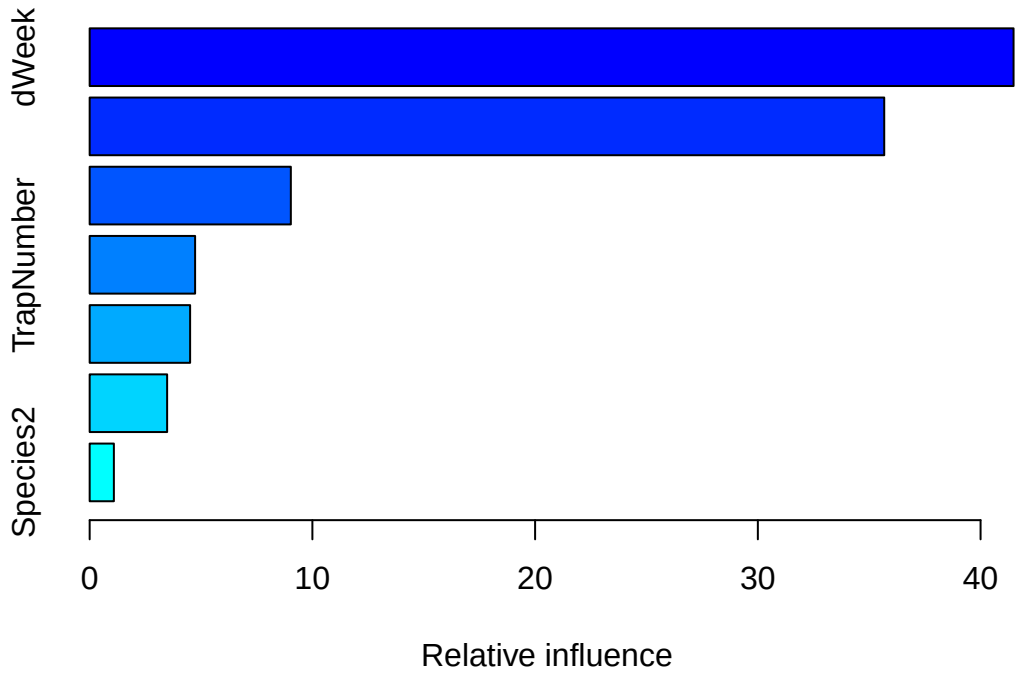
```
            cv.folds=5,shrinkage = 0.005,n.minobsinnode=10,distribution = "adaboost")
#distribution = "bernoulli"
best.iter <- gbm.perf(fitCv,method="cv")
```




```
best.iter
```

```
## [1] 1986
```

```
summary(fitCv,n.trees=best.iter)
```




```
##                         var   rel.inf
## dWeek                 dWeek 41.481649
## NumMosquitos NumMosquitos 35.675332
## dYear                 dYear  9.032732
## TrapNumber       TrapNumber  4.734668
## Longitude         Longitude  4.510609
## Latitude           Latitude  3.478160
## Species2           Species2  1.086849
```

```
p2<-predict(fitCv, newdata = xcv,n.trees = best.iter, type = "response")
auc(xcv$WnvPresent,p2)
```

```
## [1] 0.8959009
```