Untitled

AnhVu

1/9/2020

```
# Create grid table to match results from the search itself
cs <- factor(rep(c(100, 10, 1), each=12))
degs <- factor(rep(
    rep(c(15, 20, 25, 30), each=3),
  times=3))
win_ins <- factor(rep(c(60,70,80), times=12))
```

Acceptor model

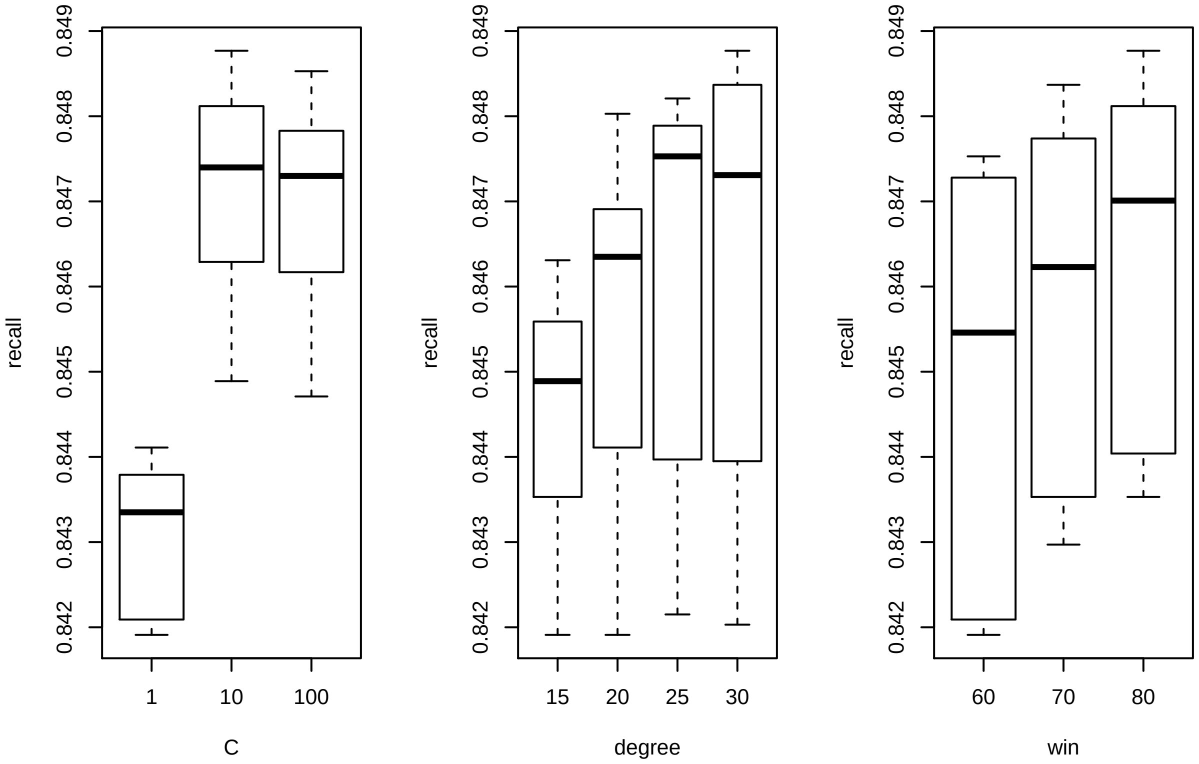
Recall

```
recalls <- read.table("data/gridsearch-acceptor-recalls.txt")
d <- recalls$V1

# Prepare table
recalls.df.flat <- data.frame(recall=d, C=cs, degree=degs, win=win_ins)
```

From the plots we see, that **recall prefers high regularization constant C (obviously) and low degree of kernels**

```
par(mfrow=c(1,3))
boxplot(recall ~ C, data=recalls.df.flat)
boxplot(recall ~ degree, data=recalls.df.flat)
boxplot(recall ~ win, data=recalls.df.flat)
```



```
aov.out <- aov(recall ~ C, data=recalls.df.flat)
summary(aov.out)
```

```
##             Df    Sum Sq   Mean Sq F value  Pr(>F)
## C            2 1.277e-04 6.384e-05   51.69 6.8e-11 ***
## Residuals   33 4.076e-05 1.240e-06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov.out <- aov(recall ~ degree, data=recalls.df.flat)
summary(aov.out)
```

```
##             Df    Sum Sq   Mean Sq F value Pr(>F)
## degree       3 1.877e-05 6.258e-06   1.338  0.279
## Residuals   32 1.497e-04 4.677e-06
```

```
aov.out <- aov(recall ~ win, data=recalls.df.flat)
summary(aov.out)
```

```
##             Df    Sum Sq   Mean Sq F value Pr(>F)
## win          2 1.348e-05 6.739e-06   1.435  0.253
## Residuals   33 1.550e-04 4.696e-06
```

To compare results for $C = 100$ and $C = 1$ we can notice, that recall for all pairs of *degree* and *window* combinations, the SVMs with $C = 1$ are worse.

```
recalls.C100 <- recalls.df.flat[which(recalls.df.flat$C == 100),]
recalls.C1 <- recalls.df.flat[which(recalls.df.flat$C == 1),]
recalls.C100$recall - recalls.C1$recall
```

```
##  [1] 0.002799104 0.001939379 0.002779111 0.004118682 0.003238964 0.003598848
##  [7] 0.005138356 0.004158669 0.003978727 0.005278311 0.005098369 0.004578535
```

Conclusion: Results for $C = 100$ and $C = 10$ are the same. $C = 1$ is worse in all aspects, regardles of *degree* and *window*. The size of window seem to increase recall, but not significantly (at the cost of higher computation time). The recall is basically influenced only by C (as is also shown by ANOVA).

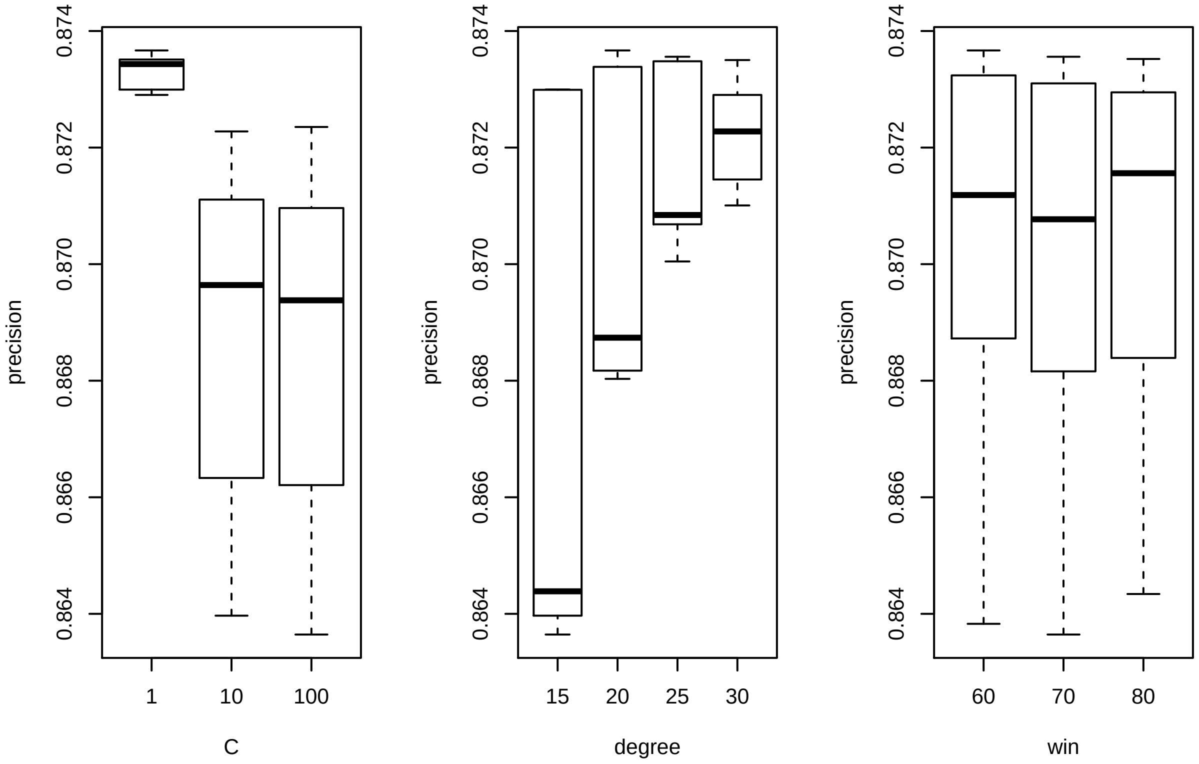
Precisions

Precision optimal parametrization goes against the optimal parametrization of recall (it prefers **low C and high degree of kernel**):

```
precisions <- read.table("data/gridsearch-acceptor-precisions.txt")
p <- precisions$V1
precisions.df.flat <- data.frame(precision=p, C=cs, degree=degs, win=win_ins)
```

Regularization constant doesn't seem to have an effect, even though we can see, that $C = 1$ is somewhat better in precision (which is in contrary to recall findings, where $C = 1$ was the worse).

```
par(mfrow = c(1,3))
boxplot(precision ~ C, data=precisions.df.flat)
boxplot(precision ~ degree, data=precisions.df.flat)
boxplot(precision ~ win, data=precisions.df.flat)
```



```
aov.out <- aov(precision ~ C, data=precisions.df.flat)
summary(aov.out)
```

```
##             Df    Sum Sq   Mean Sq F value   Pr(>F)    
## C            2 0.0001691 8.458e-05   13.53 5.11e-05 ***
## Residuals   33 0.0002063 6.250e-06                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov.out <- aov(precision ~ degree, data=precisions.df.flat)
summary(aov.out)
```

```
##              Df    Sum Sq   Mean Sq F value  Pr(>F)   
## degree        3 0.0001411 4.704e-05   6.425 0.00157 **
## Residuals    32 0.0002343 7.320e-06                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov.out <- aov(precision ~ win, data=precisions.df.flat)
summary(aov.out)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## win           2 0.0000011 5.260e-07   0.046  0.955
## Residuals    33 0.0003744 1.135e-05
```

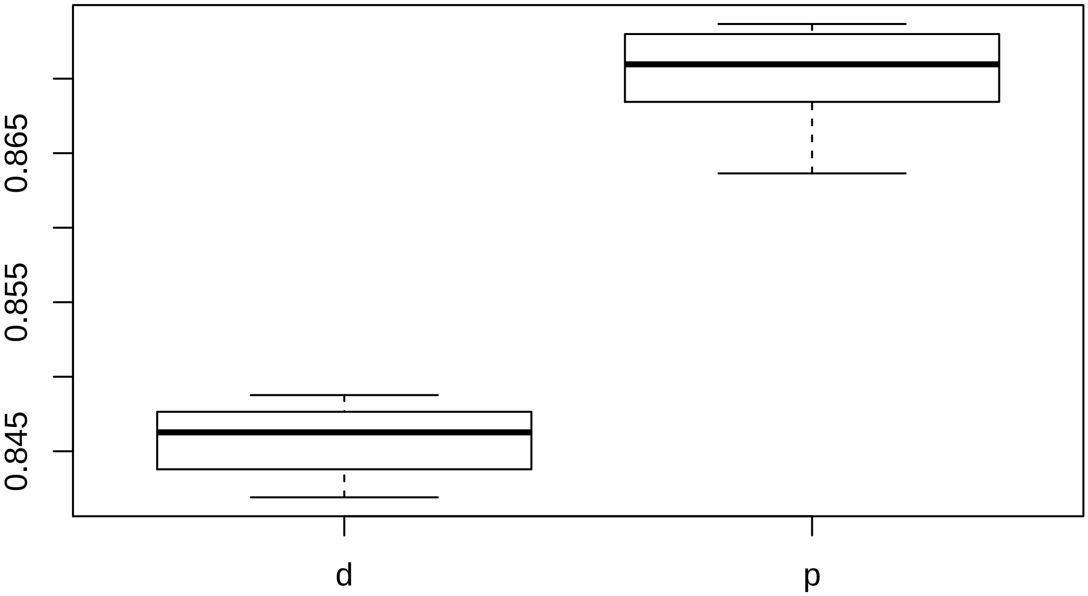
We see the effect of degree on precision is rather prominent in overall data, however it has no effect for e.g. $C = 100$

```
precisions.C100 <- precisions.df.flat[which(precisions.df.flat$C==100),]
aov.out <- aov(precision ~ win, data=precisions.C100)
summary(aov.out)
```

```
##              Df    Sum Sq  Mean Sq F value Pr(>F)
## win           2 9.900e-07 4.94e-07   0.042  0.959
## Residuals     9 1.053e-04 1.17e-05
```

Finally we see the variance in recall and precision overall:

```
boxplot(cbind(d, p))
```



The metrics seem to be equally spread. Therefore we will choose parameters to match the Ascomycota model for consistency. I.e. $C = 10$ (to emphasise recall), $win = 70$ (as window seem to slightly improve recall, no effect in precision) and $d = 25$ (prominent beneficial effect on precision, negligible benefits in recall). We can consider increasing d to 30.

Donor model

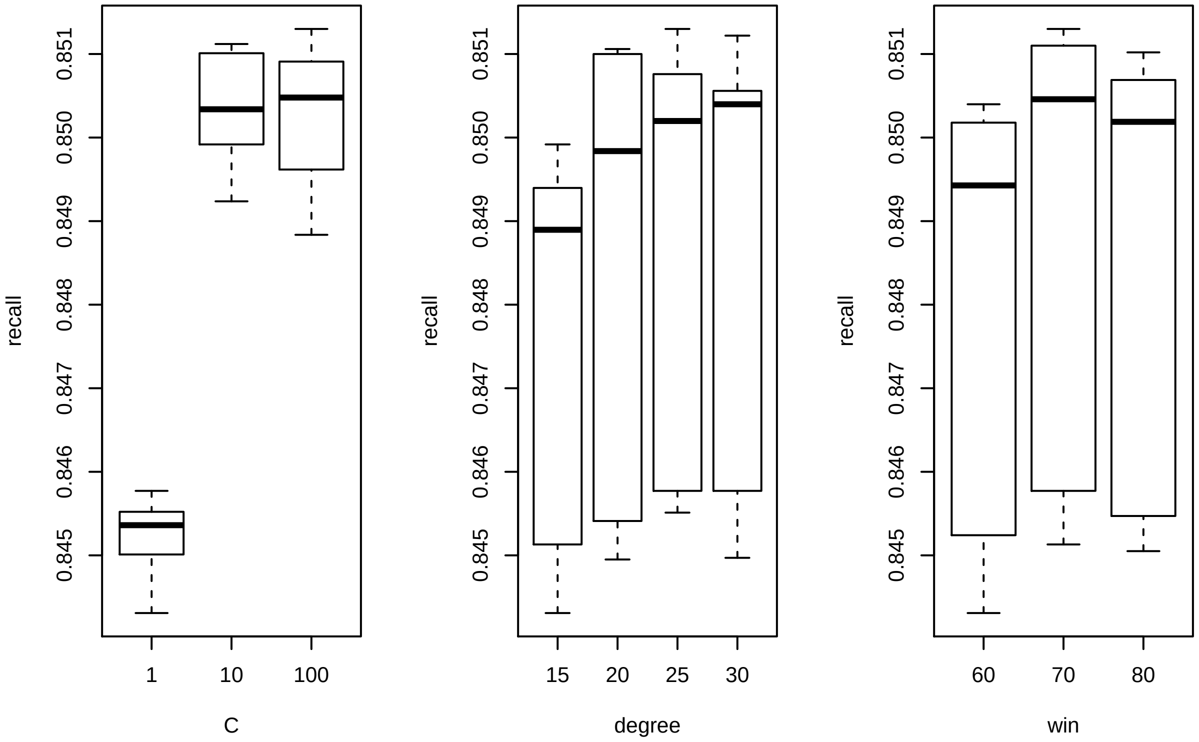
Recall

```
d_recalls <- read.table("data/gridsearch-donor-recalls.txt")
dr <- d_recalls$V1

# Prepare table
recalls.df.flat <- data.frame(recall=dr, C=cs, degree=degs, win=win_ins)
```

Similar results to acceptor site - **recall prefers high regularization constant C and low degree**.

```
par(mfrow=c(1,3))
boxplot(recall ~ C, data=recalls.df.flat)
boxplot(recall ~ degree, data=recalls.df.flat)
boxplot(recall ~ win, data=recalls.df.flat)
```



```
aov.out <- aov(recall ~ C, data=recalls.df.flat)
summary(aov.out)
```

```
##             Df    Sum Sq   Mean Sq F value Pr(>F)    
## C            2 2.045e-04 1.023e-04   238.4 <2e-16 ***
## Residuals   33 1.415e-05 4.300e-07                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov.out <- aov(recall ~ degree, data=recalls.df.flat)
summary(aov.out)
```

```
##             Df    Sum Sq   Mean Sq F value Pr(>F)
## degree       3 7.490e-06 2.498e-06   0.379  0.769
## Residuals   32 2.112e-04 6.599e-06
```

```
aov.out <- aov(recall ~ win, data=recalls.df.flat)
summary(aov.out)
```

```
##             Df    Sum Sq   Mean Sq F value Pr(>F)
## win          2 4.120e-06 2.059e-06   0.317  0.731
## Residuals   33 2.146e-04 6.502e-06
```

To compare results for $C = 100$ and $C = 1$ we can notice, that recall for all pairs of *degree* and *window* combinations, the SVMs with $C = 1$ are again consistently worse.

```
recalls.C1 <- recalls.df.flat[which(recalls.df.flat$C == 1),]
recalls.C100 <- recalls.df.flat[which(recalls.df.flat$C == 100),]
recalls.C100$recall - recalls.C1$recall
```

```
##  [1] 0.004586788 0.004266314 0.003785603 0.004887233 0.005648360 0.005207707
##  [7] 0.004646877 0.005528182 0.005227737 0.005428033 0.005448063 0.005247767
```

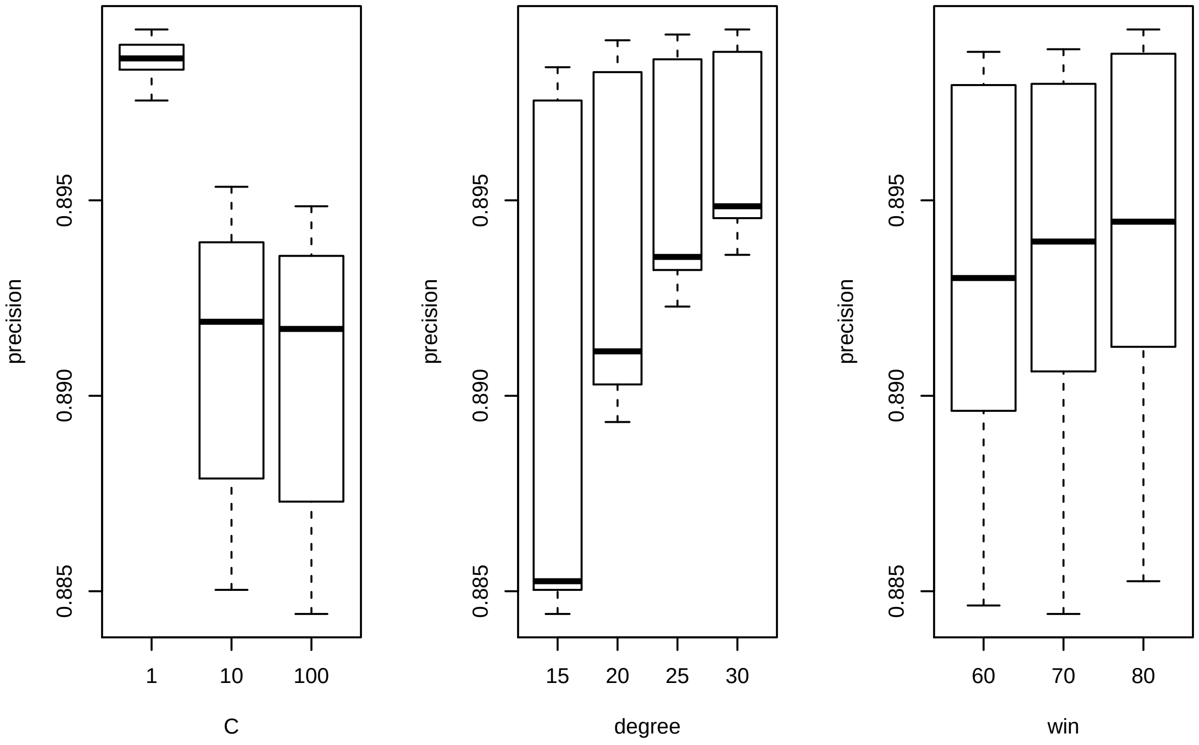
Conclusion: Results for $C = 100$ and $C = 10$ are the same. $C = 1$ is worse in all aspects, regardles of *degree* and *window*. The size of the window nor the kernel degree seem to affect recall.

Precisions

```
precisions <- read.table("data/gridsearch-donor-precisions.txt")
dp <- precisions$V1
precisions.df.flat <- data.frame(precision=dp, C=cs, degree=degs, win=win_ins)
```

Regularization constant doesn't seem to have an effect, even though we can see, that $C = 1$ is somewhat better in precision (which is in contrary to recall findings, where $C = 1$ was the worse).

```
par(mfrow = c(1,3))
boxplot(precision ~ C, data=precisions.df.flat)
boxplot(precision ~ degree, data=precisions.df.flat)
boxplot(precision ~ win, data=precisions.df.flat)
```



```
aov.out <- aov(precision ~ C, data=precisions.df.flat)
summary(aov.out)
```

```
##             Df    Sum Sq   Mean Sq F value   Pr(>F)
## C            2 0.0004851 2.425e-04   24.68 2.79e-07 ***
## Residuals   33 0.0003243 9.830e-06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov.out <- aov(precision ~ degree, data=precisions.df.flat)
summary(aov.out)
```

```
##             Df    Sum Sq   Mean Sq F value Pr(>F)
## degree       3 0.0002346 7.818e-05   4.353 0.0111 *
## Residuals   32 0.0005748 1.796e-05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov.out <- aov(precision ~ win, data=precisions.df.flat)
summary(aov.out)
```

```
##             Df    Sum Sq   Mean Sq F value Pr(>F)
## win          2 0.0000072 3.622e-06   0.149  0.862
## Residuals   33 0.0008021 2.431e-05
```

Overall, we C has the greatest effect and we again choose $C = 10$ to emphasise recall, $d = 25$ (for consistency with Ascomycota model; can be probably raised to 30 though). Window is set to 70 to emphasise recall (even though almost insignificantly)