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

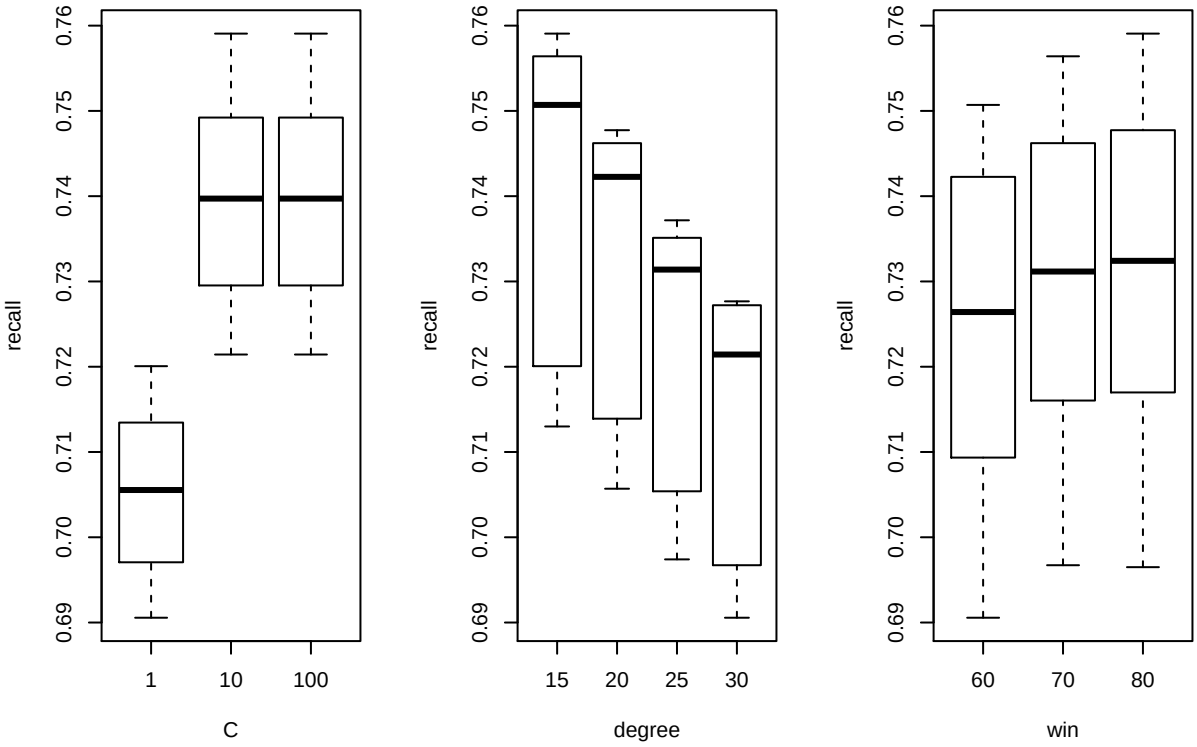
Recall

```
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 0.009191 0.004596   35.62 5.73e-09 ***
## Residuals   33 0.004258 0.000129
## ---
## 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 0.003881 0.001294   4.327 0.0114 *
## Residuals   32 0.009568 0.000299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##             Df   Sum Sq   Mean Sq F value Pr(>F)
## win          2 0.000306 0.0001528   0.384  0.684
## Residuals   33 0.013144 0.0003983
```

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] 0.037710 0.036949 0.039002 0.036569 0.033605 0.033832 0.033985 0.030791
```

```
##  [9] 0.031780 0.030867 0.030488 0.031172
```

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 does not seem to affect recall. The recall is most influenced by *degree* ($p = 2.72e - 05$ for $C = 100$), its power is however smaller for overal data ($p = 0.0114$).

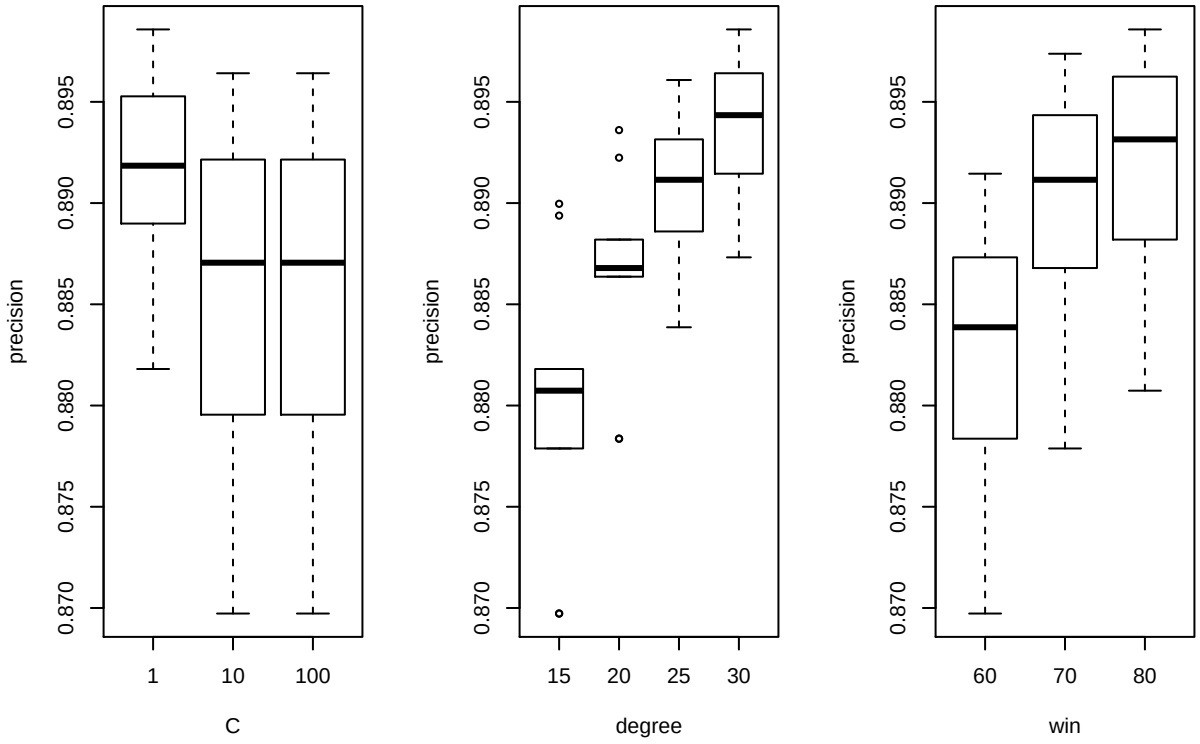
Precisions

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 0.0002878 1.439e-04   2.931 0.0674 .
## Residuals   33 0.0016205 4.911e-05
## ---
## 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 0.0009828 0.0003276   11.33 3.19e-05 ***
## Residuals   32 0.0009256 0.0000289
## ---
## 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 0.0005498 2.749e-04   6.677 0.00367 **
## Residuals   33 0.0013586 4.117e-05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

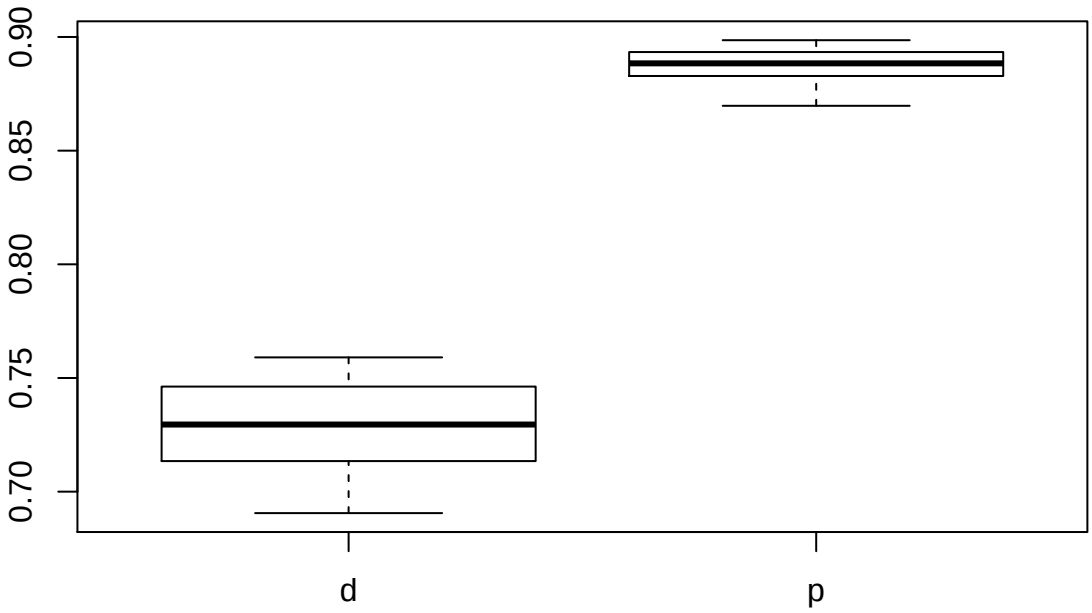
We see the effect of degree on precision is very prominent in overall data, however less so for e.g. $C = 100$

```
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 0.0002135 1.067e-04   2.046  0.185
## Residuals    9 0.0004694 5.216e-05
```

Finally we see the variance in recall and precision overall:

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



From here we decide, that we should optimize recall parameters. Even though it will adversely influence precision, the effect won't be big

Donor model

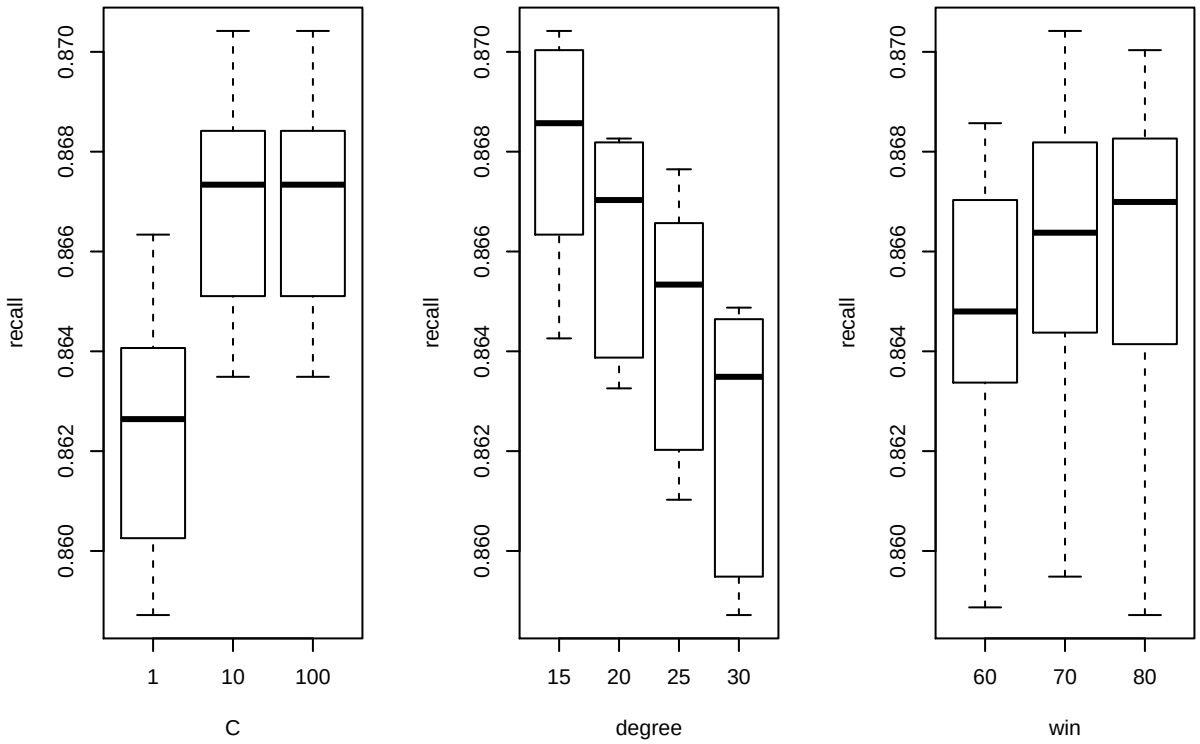
Recall

```
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 0.0001731 8.655e-05   15.95 1.42e-05 ***
## Residuals   33 0.0001790 5.430e-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 0.0001600 5.333e-05   8.881 0.000199 ***
```

```
## Residuals   32 0.0001922 6.000e-06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##             Df    Sum Sq   Mean Sq F value Pr(>F)
## win          2 0.0000117 5.833e-06   0.565  0.574
## Residuals   33 0.0003405 1.032e-05
```

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.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] 0.004312 0.004235 0.003696 0.003773 0.004312 0.004620 0.003772 0.004543
##  [9] 0.006621 0.004619 0.005389 0.005928
```

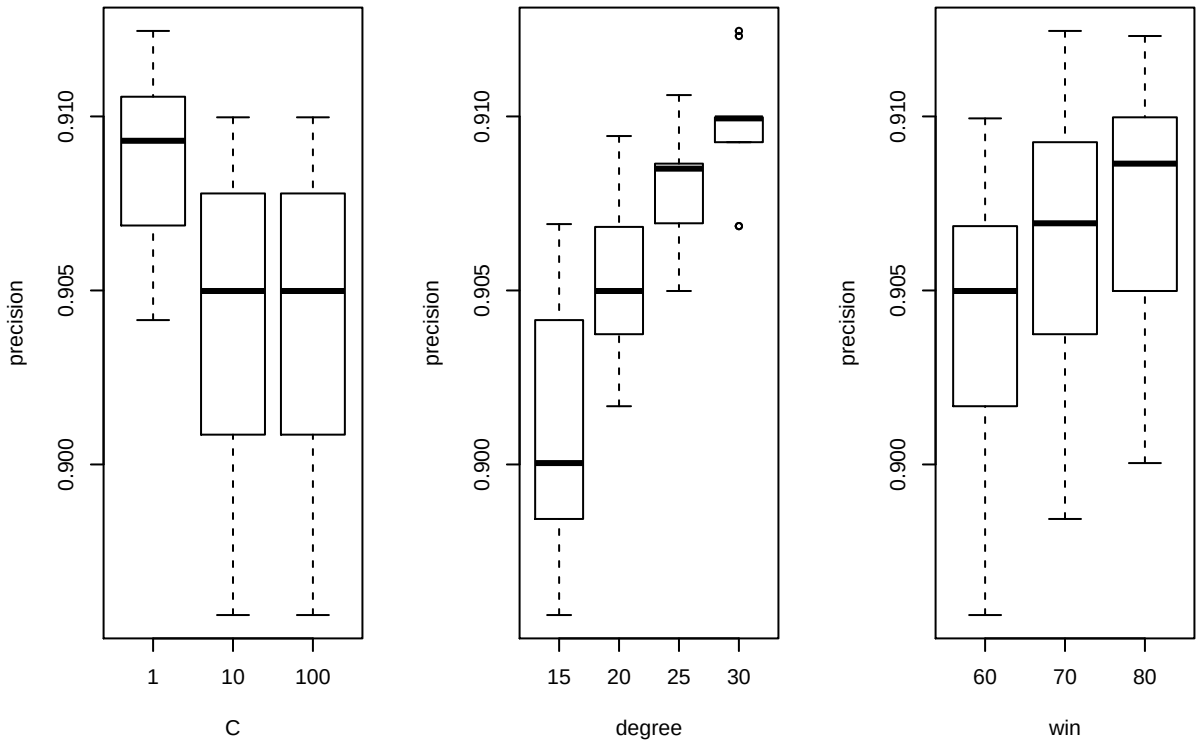
Conclusion: Results for $C = 100$ and $C = 10$ are the same. $C = 1$ is worse in all aspects, regardles of *degree* and *window*. The size of window does not seem to affect recall.

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 0.0001734 8.672e-05   5.527 0.00851 **
## Residuals   33 0.0005178 1.569e-05
## ---
## 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 0.0004173 1.391e-04   16.24 1.36e-06 ***
## Residuals   32 0.0002740 8.560e-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 0.0000660 3.299e-05   1.741  0.191
## Residuals   33 0.0006253 1.895e-05
```