Sparse vs Dense Classification

Camille

Datasets

- Breast cancer – Wisconsin – small number of features / low dimensions
  - n=699, p=8
  - Dense data – lots of data to represent each feature
  - 2 predicted classes: benign and malignant

- CAD – large number of features / high dimensions
  - n = 303, p = 54
  - More sparse data – less data to represent each feature
  - 2 predicted classes – presence or absence of coronary artery disease (CAD)
  - R. Alizadehsani et al., “A data mining approach for diagnosis of coronary artery disease,” Computer Methods and Programs in Biomedicine, vol.111, no.1, pp.52-61, Jul. 2013.

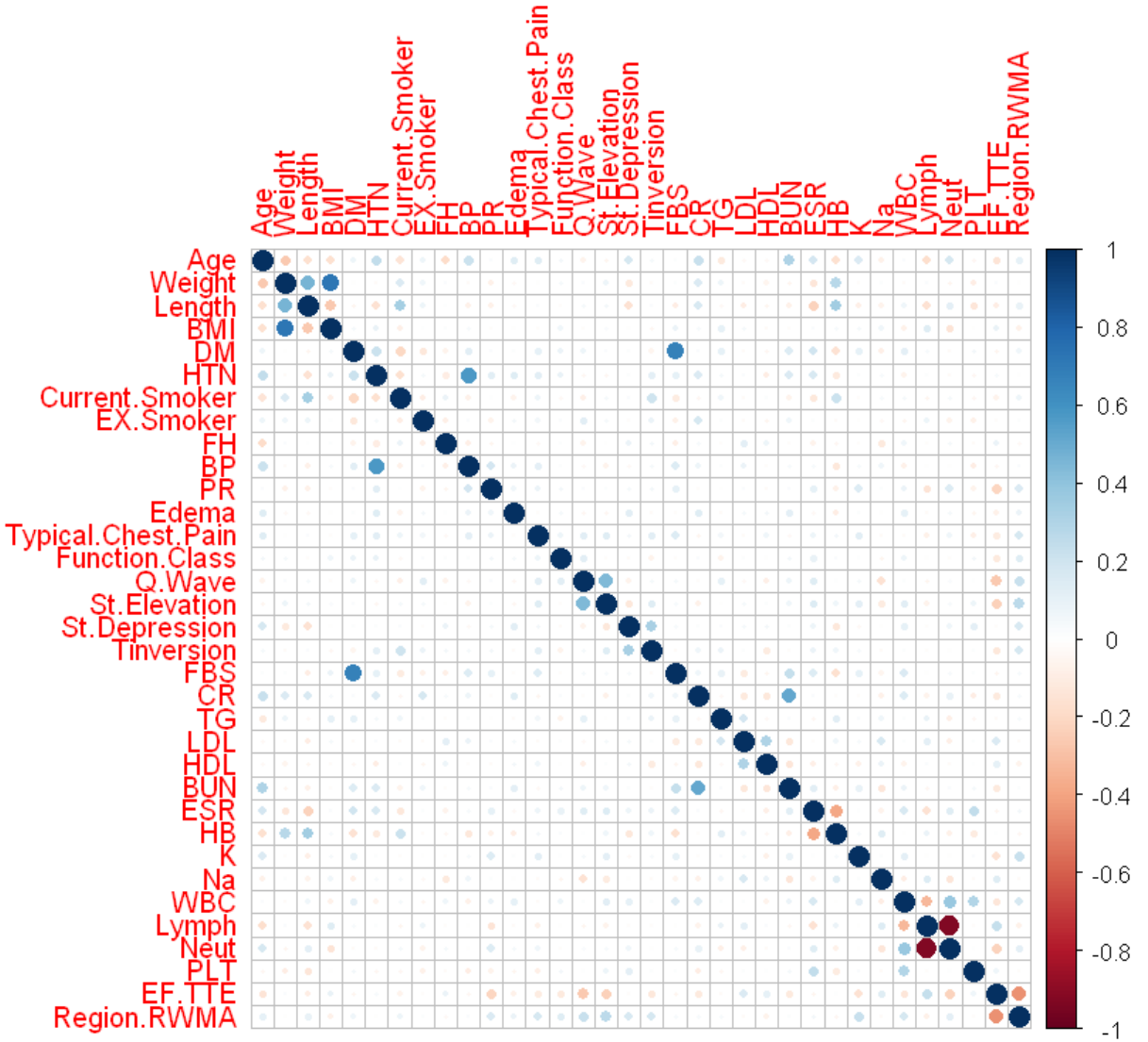
Analysis Methods / Curse of Dimensionality

- Logistic Regression and Sparse Logistic Regression
- Too many features results in overfitting the model
- Classifiers that model decision boundaries very accurately are less prone to overfitting – logistic regression will be less prone to overfitting
- When n=p or n<p logistic regression still becomes biased
- We can add a penalty to the equation similar to a lasso penalty with sparse logistic regression. This will force some less important variables to have coefficients of 0 and simplify our model.

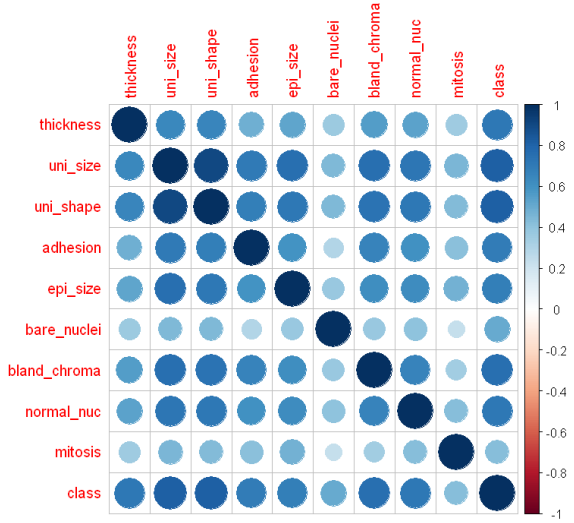
Correlation

- When there are correlated groups in the data, it is noisy data.
- We can add group lasso – where we make sure to treat groups of correlated variables in the same way – if we have highly correlated data.

CAD data



Breast Cancer Data



- The variables are only a little correlated for the CAD data, but they are highly correlated for all of the breast cancer data
- Group lasso will not be needed for the CAD data

Logistic Regression with Cancer Data

- After running logistic regression, I can see that not all the variables have significant p-values. Thickness, adhesion, bare nuclei, and bland chromatin are very significant.
- When I run ANOVA, every variable significantly decreases the residual deviance, so all should be added to the final model.
- With a threshold of .5
- Sensitivity = 96.9%
- Specificity = 93.3%
- The model is performing well.

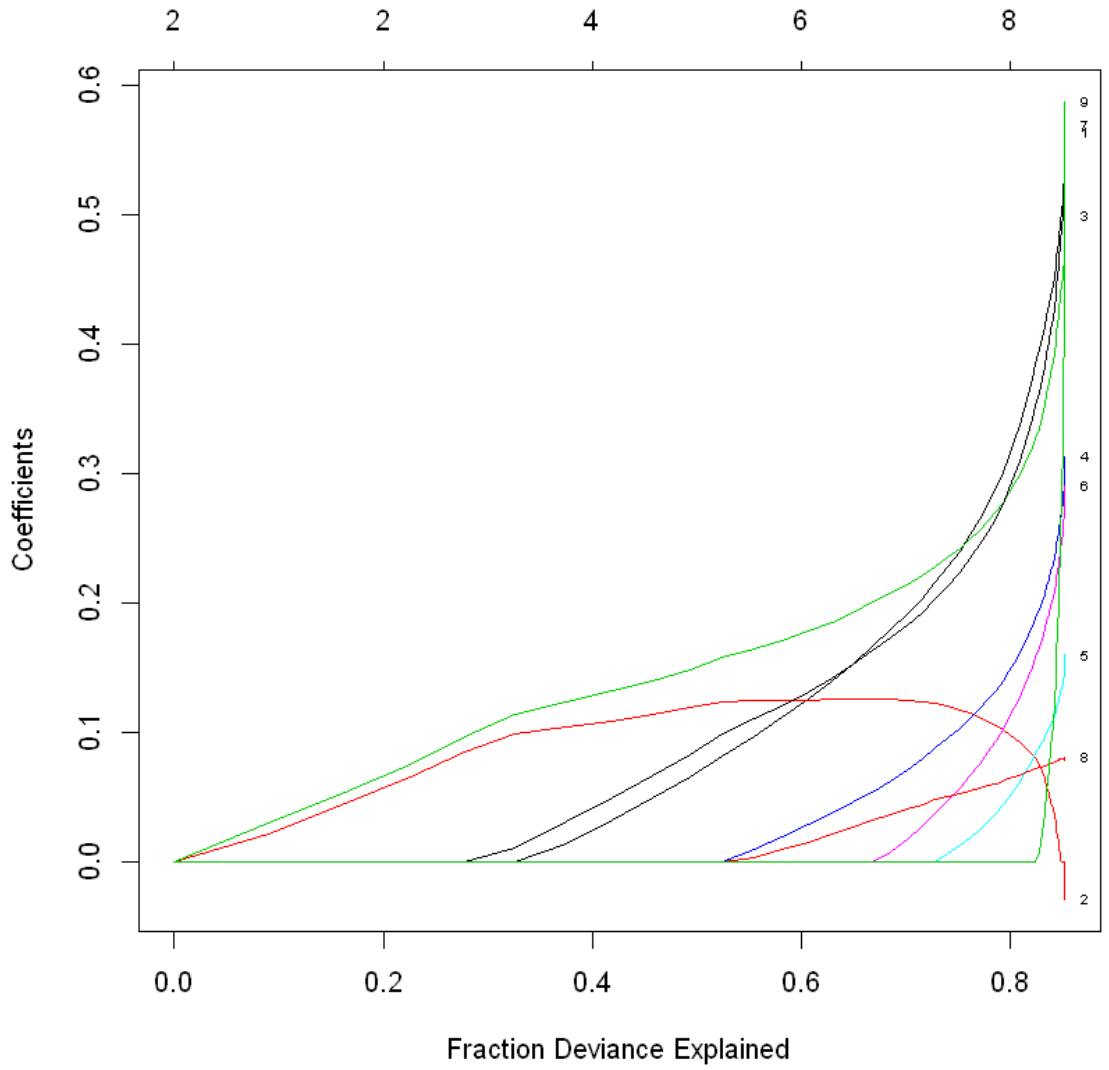
```
Call:
glm(formula = class ~ thickness + uni_size + adhesion + epi_size +
    bare_nuclei + bland_chroma + normal_nuc + mitosis, family = binomial(link = "logit"),
    data = breast_cancer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7560  -0.1332  -0.0714   0.0257   2.3586

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.83531    1.14555  -9.459  < 2e-16 ***
thickness     0.62150    0.11856   5.242 1.59e-07 ***
uni_size      0.28410    0.15031   1.890 0.058745 .
adhesion      0.32820    0.09665   3.396 0.000684 ***
epi_size      0.22328    0.13386   1.668 0.095302 .
bare_nuclei   0.31387    0.10672   2.941 0.003271 **
bland_chroma  0.62800    0.14987   4.190 2.79e-05 ***
normal_nuc    0.10858    0.09702   1.119 0.263115
mitosis       0.57817    0.31931   1.811 0.070187 .
```

ANOVA

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	698	900.5274	NA
thickness	1	436.473481	697	464.0540	6.338108e-97
uni_size	1	251.747248	696	212.3067	1.080262e-56
adhesion	1	22.413970	695	189.8927	2.197697e-06
epi_size	1	9.437713	694	180.4550	2.125684e-03
bare_nuclei	1	15.481715	693	164.9733	8.330725e-05
bland_chroma	1	20.770283	692	144.2030	5.178023e-06
normal_nuc	1	1.523235	691	142.6798	2.171307e-01
mitosis	1	3.606688	690	139.0731	5.754763e-02

Sparse Logistic Regression with Cancer Data

- I chose misclassification error as the optimizer for lambda
- All of the covariates are included in the final model
- Sensitivity = 97.4%
- Specificity = 94.6%
- The final model is the same and the sensitivity / specificity are about the same for logistic regression and sparse logistic regression.
- For this dataset, since all covariates are included with both methods, logistic regression is better and simpler.



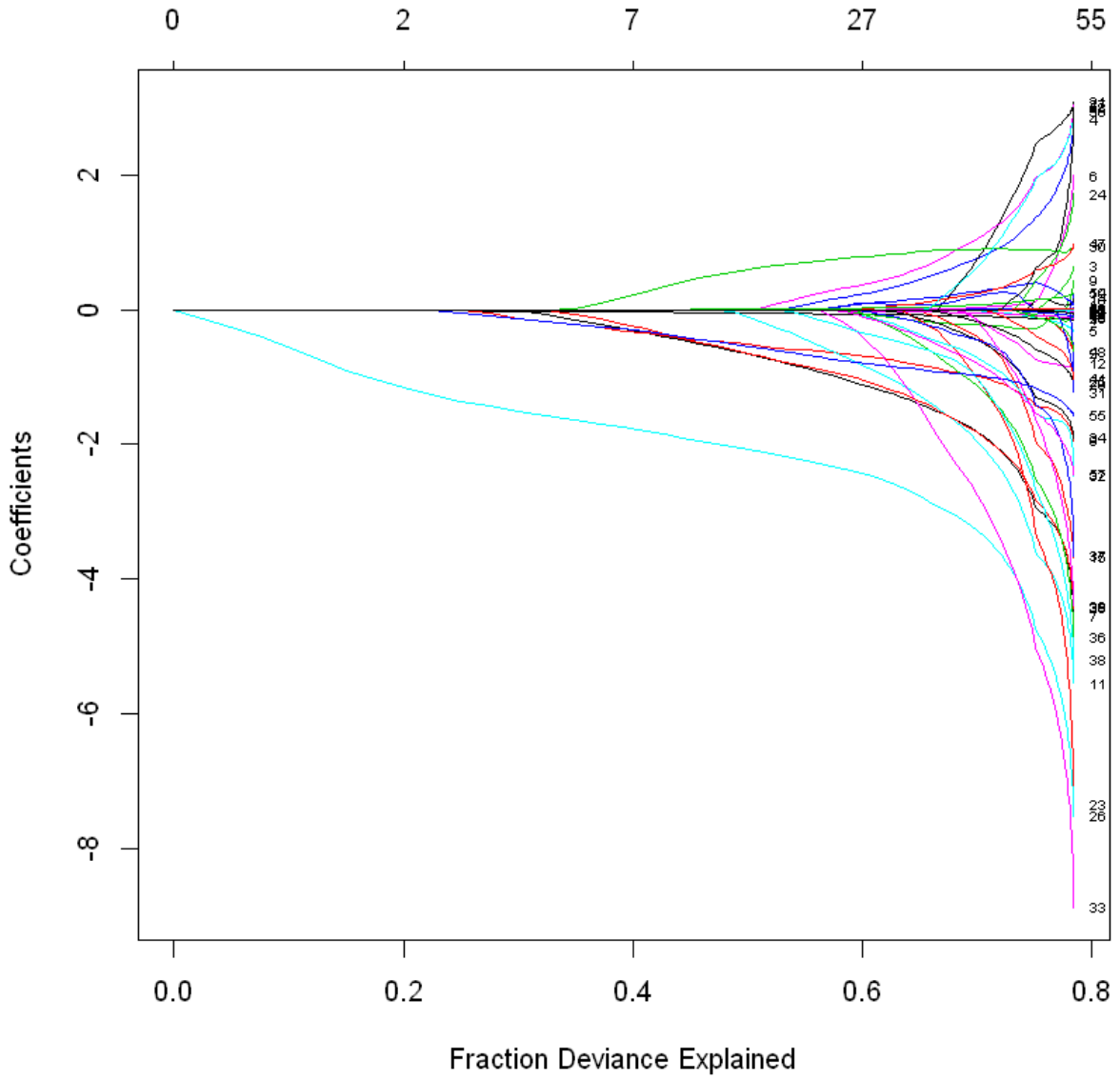
Logistic Regression with CAD Data

- Only a subset of the results are posted.
- There are some variables that are significant, but not all
- When I run ANOVA, the results are that all 54 variables should be included.
- Sensitivity = 95.8%,
- Specificity = 93%
- Reasonable rates, difficult to interpret model with so many covariates
- It could be more useful to know a few variables that are predictive
- Sparse logistic regression can make a model with less covariates

```
glm(formula = Cath ~ Age + Weight + Length + Sex + BMI + DM +
    HTN + Current.Smoker + EX.Smoker + FH + Obesity + CRF + CVA +
    Airway.disease + Thyroid.Disease + CHF + DLP + BP + PR +
    Edema + Weak.Peripheral.Pulse + Lung.rales + Systolic.Murmur +
    Diastolic.Murmur + Typical.Chest.Pain + Dyspnea + Function.Class +
    Atypical + Nonanginal + LowTH.Ang + Q.Wave + St.Elevation +
    St.Depression + Tinversion + LVH + Poor.R.Progression + BBB +
    FBS + CR + TG + LDL + HDL + BUN + ESR + HB + K + Na + WBC +
    Lymph + Neut + PLT + EF.TTE + Region.RWMA + VHD, family = binomial(link = "logit"),
    data = CAD)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1941  -0.0191  -0.0002   0.0078   2.6880

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -8.599e+01  5.480e+01  -1.569  0.11657
Age                 -1.609e-01  5.278e-02  -3.048  0.00230 **
Weight              -7.666e-01  3.030e-01  -2.530  0.01140 *
Length               7.497e-01  2.976e-01   2.519  0.01177 *
SexMale             -3.443e+00  1.867e+00  -1.844  0.06522 .
BMI                  2.297e+00  8.635e-01   2.659  0.00783 **
DM                  -4.987e+00  2.112e+00  -2.361  0.01823 *
HTN                 -2.089e+00  1.646e+00  -1.269  0.20431
Current.Smoker       6.042e-01  1.265e+00   0.477  0.63304
EX.Smoker            4.240e-01  6.647e+00   0.064  0.94914
```

Sparse Logistic Regression with CAD Data

- I chose misclassification error as the optimizer for lambda
- I didn't use group lasso
- The final model includes 24 out of 54 covariates, which is much easier to interpret and use
- Sensitivity = 94.9%
- Specificity = 79.3%
- The specificity of the model is lower, but this is not as important in a medical setting as sensitivity, which is about the same. This model may still be useful.



Discussion

- The CAD dataset had more sparse data than the cancer dataset, but not enough to have problems with regular logistic regression
- Finding a model that only needs a small number of covariates is more useful for the medical setting. It is costly and time consuming to measure things, so finding what is most useful to measure to predict an outcome is good.
- A dataset where n = p or n < p would likely have more problems with accuracy and be more in need of special methods like sparse logistic regression

Conclusions

- Sparse Logistic Regression is useful for data where n = p or n < p
- It can also be useful if we have lots of parameters in our data and we want to have less in our final model
- For datasets where n is much greater than p, regular logistic regression is probably better
- There are also other methods that can be useful for highly correlated data like group lasso