Compare Linear Classifiers

LDA

Estimate **mu1**, **mu0**, **Sigma** then we have P(y) and P(x | y).

For binary classification, decision boundary is determined by

P(Y=1 | x)/P(Y=0 | x) > 0,

which corresponds to the following linear function

$$\mathbf{x}^t \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \alpha_0$$

Logistic Regression

Directly estimate log of the following ratio by a linear function, without any assumptions on P(x)

P(Y=1 | x)/P(Y=0 | x) > 0.

So if data are generated from a mixture of two normals following the assumption of LDA, then LDA and Logistic should return the same linear function, of course, asymptotically. Estimates from finites samples may differ.

Linear SVM

$$\arg\min_f \mathbb{E}\Big[1 - Y f(X)\Big]_+$$

$$= \operatorname{sign}\Big(\eta(x) - \frac{1}{2}\Big)$$

Compare Linear Classifiers

$$\eta(x) = P(Y = 1|X = x) \qquad f(x) = a =?$$

$$\mathbb{E}_{Y|x}\big[1 - Y \cdot a\big]_+ = \big[1 - a\big]_+ \cdot \eta(x) + \big[1 + a\big]_+ \cdot \big(1 - \eta(x)\big)$$

$$a \in [-1, 1]$$

$$\big[1 - a\big] \cdot \eta(x) + \big[1 + a\big] \cdot \big(1 - \eta(x)\big)$$
$$= 1 - a \cdot \big(1 - 2\eta(x)\big)$$

Linear SVM

$$\arg\min_f \mathbb{E}\Big[1 - Yf(X)\Big]_+$$
$$= \operatorname{sign}\Big(\eta(x) - \frac{1}{2}\Big)$$

Imbalanced/Unbalanced Data

Data: many Y=0, very few Y=1

Problem: cannot beat the rule that predicts everything to be class 0

1. If mis-classification rate is the goal, then go with that one-class classification rule.

2. If the two errors, classifying Y=1 to be 0 or classifying Y=0 to be 1, have different consequences, then use **asymmetric classification error**, which will lead to a prob cut-off value different from the usual 0.5.

3. Use other loss functions (for evaluation) that fit the underlying application. For example, if ranking is of interest, use **AUC**.

4. **Down-sampling** or **up-sampling** or **re-weighting**. Don't forget to **calibrate** your model at the end.
    1. Platt's scaling
    2. Isotonic Regression

Infinitely Imbalanced Logistic Regression

(by Art Owen, see paper link on Piazza)

Suppose data are

y=1, x_{1i}, i=1, …, n_1

y=0, x_{0i}, i=1, …., N

Fit a logistic regression model on this data, then consider an extreme situation, N —> infinity. Do the logistic coefficients have a meaningful limit?

Suppose

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_{1i} \in \mathbb{R}^d \quad \& \quad x \sim F_0 \quad \text{when} \quad Y = 0$$

Let $\alpha(N)$ and $\beta(N)$ be logistic regression estimates

The intercept —> (-infinity)

The slope beta converges to the following, under some mild conditions on F0

We have

$$\bar{x} = \frac{\int x\, e^{x'\beta}\, dF_0(x)}{\int e^{x'\beta}\, dF_0(x)}$$

β is the *exponential tilt* to take $E_{F_0}(X)$ onto $\bar{x}$

To understand the exponential tilt, assume LDA assumption holds.

```
> head(heart)
  sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
1 160   12.00 5.73     23.11 Present    49   25.30   97.20  52   1
2 144    0.01 4.41     28.61  Absent    55   28.87    2.06  63   1
3 118    0.08 3.48     32.28 Present    52   29.14    3.81  46   0
4 170    7.50 6.41     38.03 Present    51   31.99   24.26  58   1
5 134   13.60 3.50     27.78 Present    60   25.99   57.34  49   1
6 132    6.20 6.47     36.21 Present    62   30.77   14.14  45   0
> heart$famhist = as.numeric(heart$famhist)
> heartfull = glm(chd ~., data=heart, family=binomial)
```

```
> table(heart$chd)

  0   1
302 160
```

```
>
> id = which(heart$chd ==1)
> one.sample = apply(data.matrix(heart[id, ]), 2, mean)
> one.sample[10] = 1
> round(one.sample, dig=4)
      sbp   tobacco       ldl adiposity   famhist     typea   obesity   alcohol
 143.7375    5.5249    5.4879   28.1202    1.6000   54.4937   26.6229   19.1453
      age       chd
  50.2938    1.0000
> newheart = rbind(heart[-id, ], one.sample)
```

```
> table(newheart$chd)

  0   1
302   1
```

```
> newfit = glm(chd ~., data=newheart, family=binomial)
> round(cbind(coef(heartfull), coef(newfit)), dig=4)
               [,1]     [,2]
(Intercept) -7.0761 -12.7297
sbp          0.0065   0.0119
tobacco      0.0794   0.0872
ldl          0.1739   0.2022
adiposity    0.0186   0.0323
famhist      0.9254   1.1918
typea        0.0396   0.0394
obesity     -0.0629  -0.1015
alcohol      0.0001   0.0001
age          0.0452   0.0428
```