From testing to classification

Monday, April 24, 2017 2:18 PM

Recall classification: $Y \in \{0,1\}$

So prediction is given $X = x$ output $\hat{y} \in \{0,1\}$.

Oracle Suppose some "oracle" gave us

$$f_1(x) = f_{X|Y}(x|Y=1), \quad f_0(x) = f_{X|Y}(x|Y=0)$$

then prediction is a simple vs. simple hypothesis test: $H_0: X \sim f_0(x)$ (null) vs. $H_1: X \sim f_1(x)$ (alternative).

ex A social network company wants to determine if users i, j would become "friends". To do this they calculate the number of common friends $N_{i,j}$.

They also estimate that $f_k(N_{i,j}) \propto \exp(-\eta_k N_{i,j})$ for η_k known (previously estimated) with $\eta_1 < \eta_0$.

Likelihood ratio: $$\frac{f_1(N_{i,j})}{f_0(N_{i,j})} \propto \exp\left((\eta_0 - \eta_1) N_{i,j}\right)$$

Neyman-Pearson Testing:

$$\hat{y} = \begin{cases} 1, & (f_1/f_0)(N_{i,j}) > \tau \\ 0, & \end{cases}$$

$$y = \begin{cases} 1, & (t'/t_0)(N_{i,j}) > \tau \\ 0, & (t'/t_0)(N_{i,j}) \le \tau \end{cases}$$

$$\equiv \hat{y} = \begin{cases} 1, & N_{i,j} > K \\ 0, & N_{i,j} \le K \end{cases} \quad \text{for some } \tau, K.$$

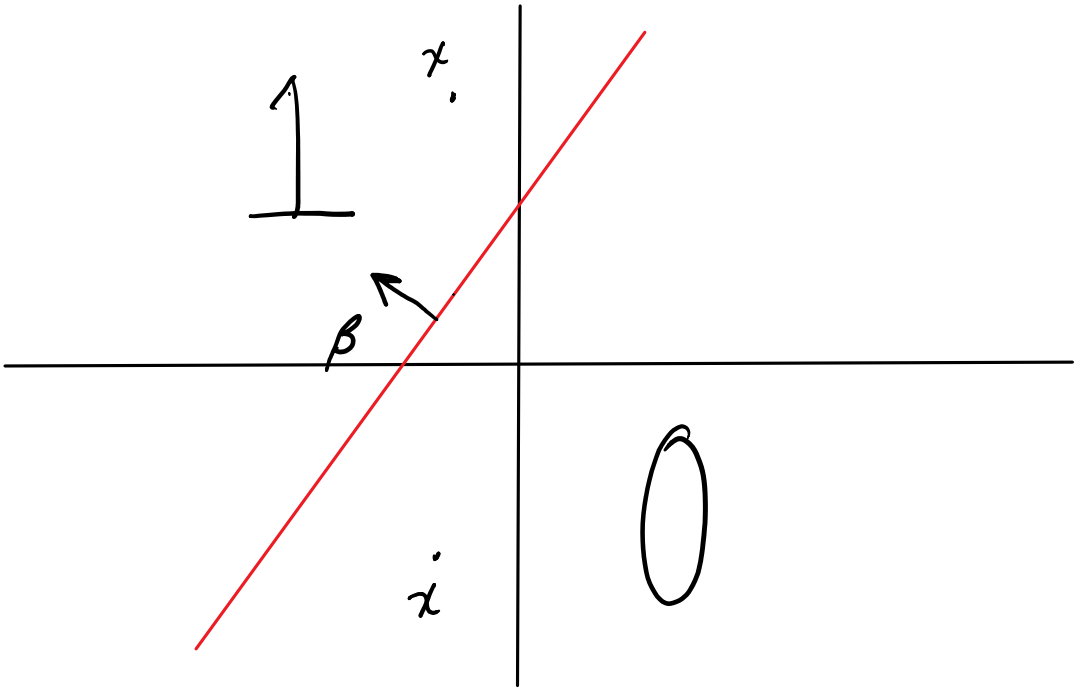
Why only use $N_{i,j}$? What about distance between i,j in shortest path length distance?

Let $x \in \mathbb{R}^p$ be p predictor variables then given weight vector $\beta \in \mathbb{R}^p$ let <u>score</u> be $\beta^T x$.

<u>ex</u>. $S_{i,j} = N_{i,j} + 5 \cdot \frac{1}{d(i,j)} = 1 \cdot x_1 + 5 \cdot x_2$.

note: $\beta^T x \ge K \equiv \beta^T x + \beta_0 \ge 0 \quad \beta_0 = -K$

so can remove K by adding intercept $x_0 = \underline{1}$.


x
1
β
0
x'

Evaluating predict

Monday, April 24, 2017 3:25 PM

Have test set $\{x_i, y_i\}$, want to evaluate predict method $f(x) = \beta^T x$.

Confusion Matrix

	Predict 1	Predict 0
Actual 1	True Positive (TP)	False Negative (FN)
Actual 0	False Positive (FP)	True Negative (TN)

Predict

1. Calculate score $s_i = f(x_i)$ for each i.
2. Order scores: $s_{a_1} \geq s_{a_2} \geq s_{a_3} \geq \dots \geq s_{a_n}$
   where $a_1, \dots, a_n$ is permutation of $\{1, \dots, n\}$.
3. Predict $\hat{y}_{a_1}, \dots, \hat{y}_{a_T} = 1$ and $\hat{y}_{a_{T+1}}, \dots, \hat{y}_{a_n} = 0$.

s:	3.1	3	2.8	2.6	2.3	2.1	1.1	1	.8	.6	.5	.2	
$\hat{y}$	1	1	1	1	1	1	0	0	0	0	0	0	$T=6$
y	1	0	1	1	0	1	0	1	0	0	1	0	

Conf. TP = 4 | FN = 2

Conf. Matrix

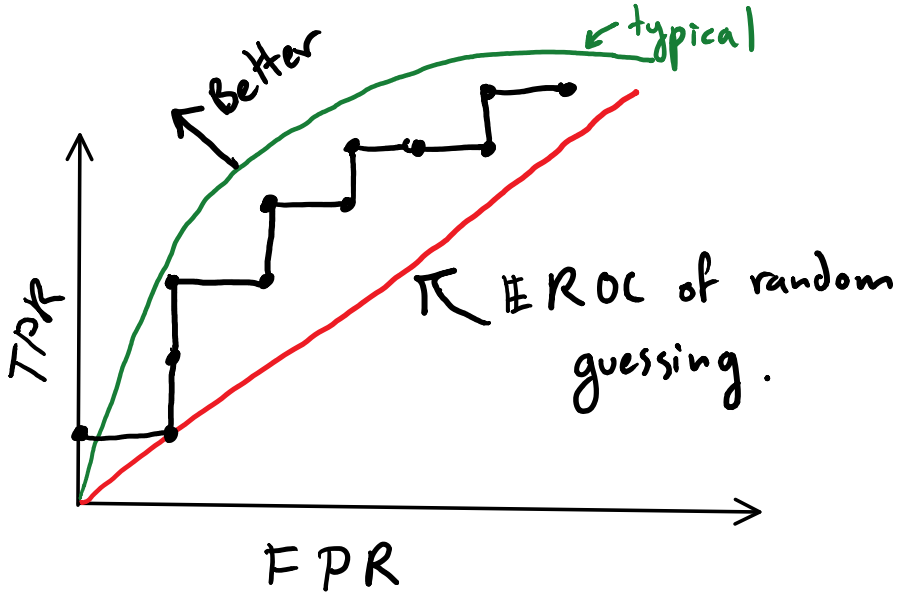
TP = 4	FN = 2
FP = 2	TN = 4

Reciever - Operating Characteristic (ROC)

True positive rate (TPR) = $\frac{TP}{TP + FN}$ } Actual Positive
(aka recall)

False positive rate (FPR) = $\frac{FP}{FP + TN}$ } Actual Neg.

T	1	2	3	4	5	6	7	8
TPR	1/6	1/6	2/6	3/6	3/6	4/6	4/6	5/6
FPR	0/6	1/6	1/6	1/6	2/6	2/6	3/6	3/6



- Every element of confusion matrix is represented in ROC

Precision - Recall Curve

In recommendation systems, many 0's – few 1's.

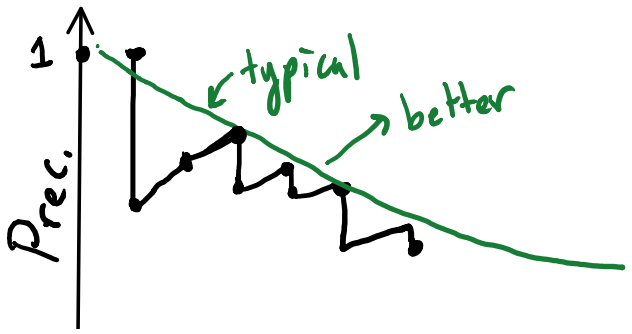
ex
- Link prediction in sparse graphs (friend recomm.)
- Search engine (most pages are not relevant)
- Document retrieval

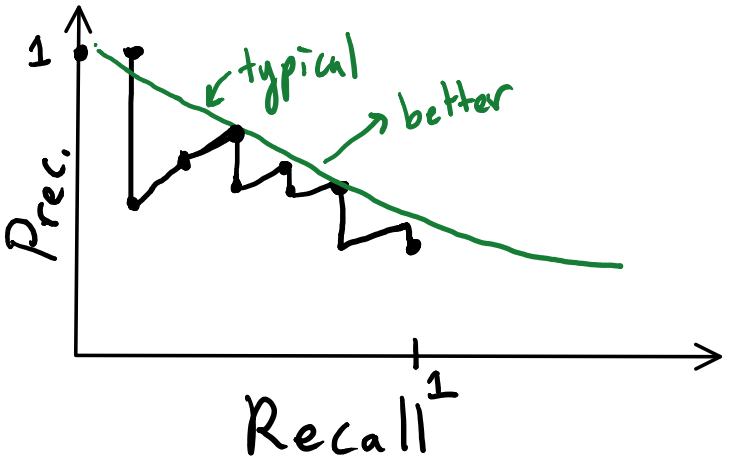
$$\text{Recall} = \text{TPR} = \frac{TP}{TP + FN}$$

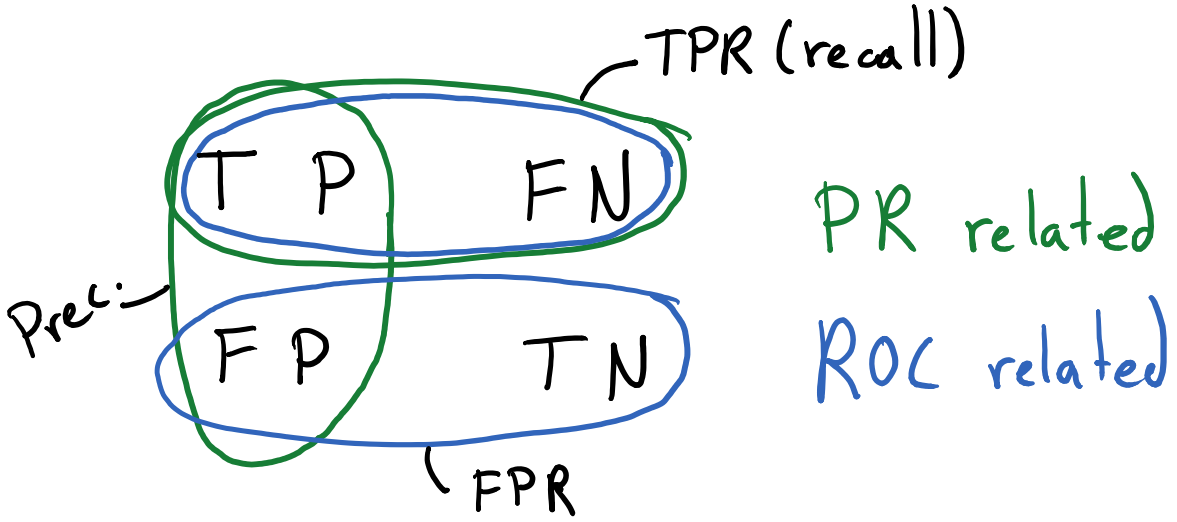
$$\text{Precision} = \frac{TP}{TP + FP}$$

} Neither is sensative to increasing TN

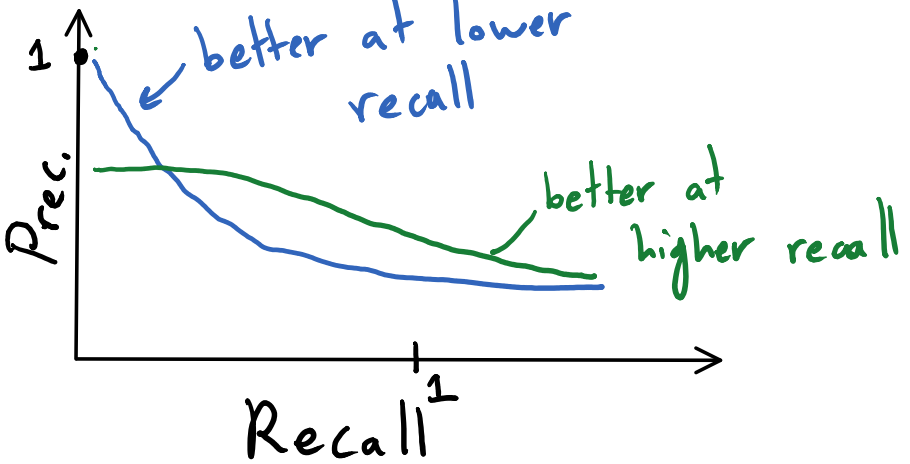
T	1	2	3	4	5	6	7	8	
Recall	1/6	1/6	2/6	3/6	3/6	4/6	4/6	5/6	...
Prec.	1	1/2	2/3	3/4	3/5	4/6	4/7	5/8	...







Comparing PR Curves



Generative Methods for Classification

Monday, April 24, 2017 5:40 PM

Ideal prediction method:

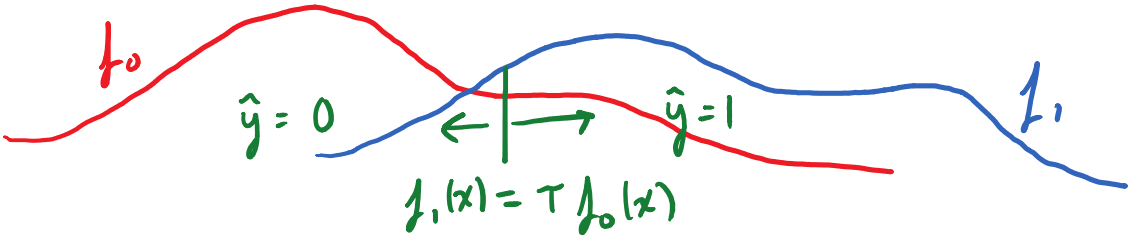
$$\hat{y} = \begin{cases} 1, & (f_1/f_0)(x) > \tau \\ 0, & \text{otherwise} \end{cases}$$

but don't have f_1, f_0.

Generative methods: Fit f_1, f_0 from training set.

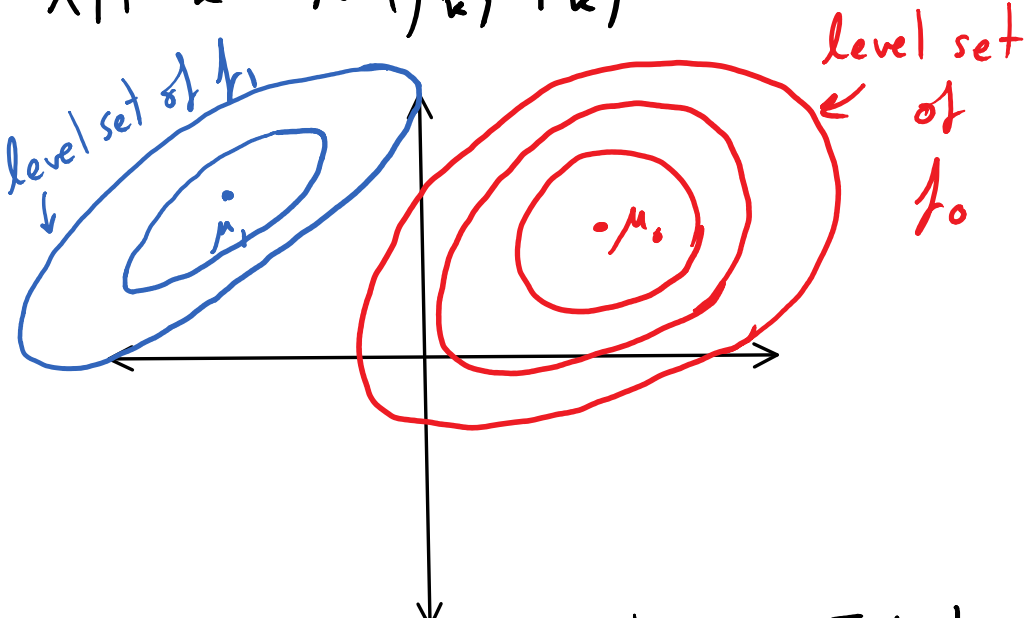
Discriminative methods: Focus on boundary

$$f_1(x) = \tau f_0(x)$$



Gaussian Generative Models

$$X|Y=k \sim N(\mu_k, \Sigma_k)$$



$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}$$

level set : $(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k) = K$.

$$\frac{f_1(x)}{f_0(x)} = \frac{|\Sigma_0|^{1/2}}{|\Sigma_1|^{1/2}} \exp\left\{-\frac{1}{2}\left[(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T \Sigma_0 (x-\mu_0)\right]\right\}$$

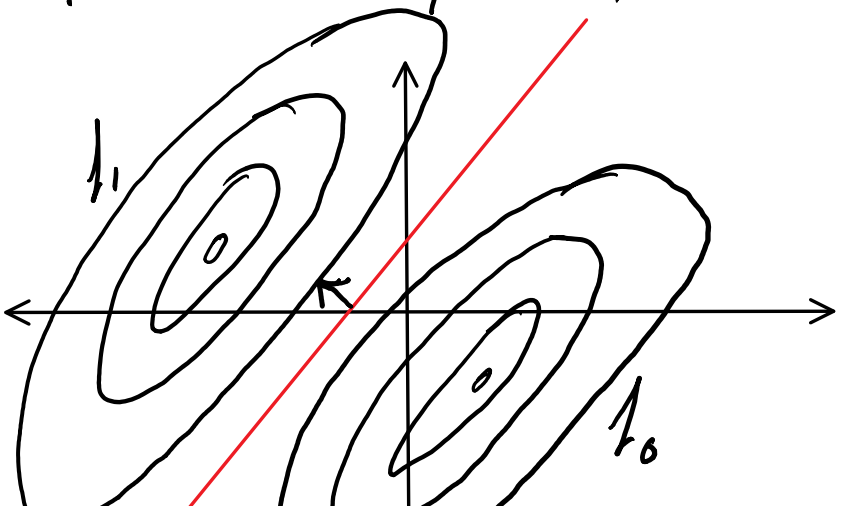
Linear Discriminant Analysis

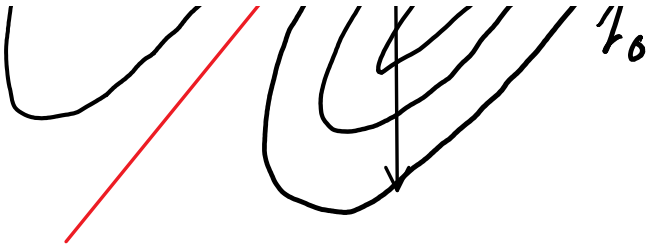
Assume that $\Sigma_1 = \Sigma_2 = \Sigma$. Likelihood ratio classifier is equivalent to

$$1\{(x-\mu_1)^T \Sigma (x-\mu_1) - (x-\mu_2)^T \Sigma (x-\mu_2) < K\}$$

$$= 1\{x^T \Sigma \mu_1 - x^T \Sigma \mu_2 > K'\} \quad \text{for some } K'$$

$$= 1\{x^T \beta > K'\} \quad \text{for } \beta = \Sigma(\mu_1 - \mu_2)$$





Assuming $\Sigma_1 = \Sigma_2$ gives a linear decision boundary.

Fitting LDA Let $\Omega_k := \{i : y_i = k\}$ then

$$\bar{x}_k = \frac{1}{|\Omega_k|} \sum_{i \in \Omega_k} x_k$$

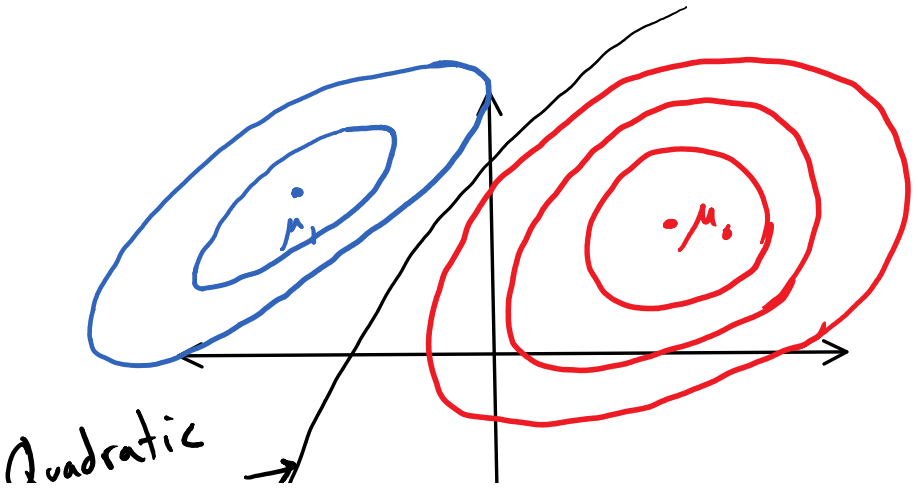
$$\hat{\Sigma} = \frac{1}{n-2} \sum_k \sum_{i \in \Omega_k} (x_k - \bar{x}_k)(x_k - \bar{x}_k)^T$$

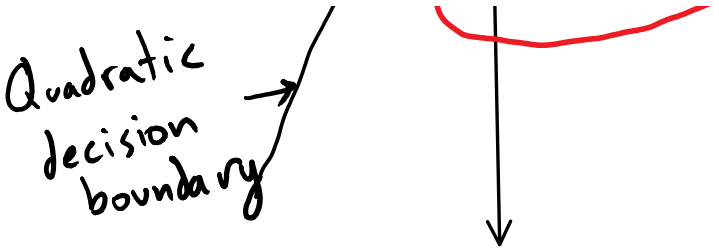
$$\hat{\beta} = \hat{\Sigma}^{-1}(\bar{x}_1 - \bar{x}_0)$$

- only fits linear boundaries
- $\hat{\Sigma}$ not invertible in high-dimensions

Quadratic Discriminant Analysis $\Sigma_1 \neq \Sigma_2$

$$\hat{y} = 1\left\{(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0) < K\right\}$$





Fit Σ_k individually ...

$$\hat{\Sigma}_k = \frac{1}{|\Omega_k| - 1} \sum_{i \in \Omega_k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^T$$

- fit more parameters (higher dimensions)
- get a quadratic boundary
- $\hat{\Sigma}_k$ not invertible in high-dimensions.

High-dimensions

Two types of solutions for high-dimensionality:

(1) Assume structure / sparsity for μ_k.

(2) Assume sparsity for Σ_k, Σ_k^{-1}.