Wednesday, September 10, 2014 3:18 PM

Notation

D : data $\{(x_i, y_i)\}_{i=1}^{N}$

D_i : data sample (x_i, y_i)

~~D~~ dimension d

θ parameter of Bernoulli $Ber(y \mid \theta)$

θ hypothesis for MLE $h \in H$ in bernoulli $\theta \in [0, 1]$

X, x_i : design matrix, features of example 'i' ~~A, a_i~~

Y, y_i : target vector, label of example 'i' ~~B, b_i~~

Naive Bayes

Issues
- div by 0 ✓ model makes no sense
- MLE?

Advantages
- simple
- ~~fast~~ train
- parallel
- small # params
- not much data needed

Disadvantages
- strong mutual dependence assumption
- ∴ limited modelling power only certain data sets

Project idea
- explore ways to relax NB assumption (TAN Bayes, Bayesian Network Classifiers) tree augmented

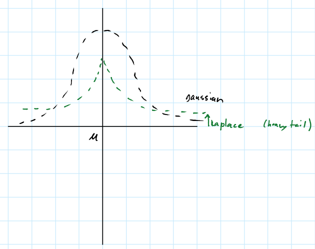
Today what if $x_i \in \mathbb{R}^D$ real valued features instead of binary

$\bar{x}_i \in \mathbb{R}^d$

Guassian Distribution

$x \sim N(\mu, \sigma^2)$ (distributed, normal)

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$



Motivation
- central limit theorum (converge to mean)
- analytic properties (smooth, differentiable, symetric)
- defined $(-\infty, \infty)$
- easy to sample
- data is Gaussian

MLE

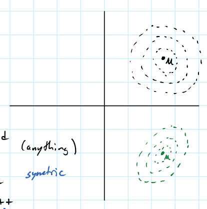
$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 \qquad \Rightarrow p(x^i \mid y_i)$$

$\bar{x} \sim N(\mu, \Sigma)$

$$p(\bar{x} \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\, |\Sigma|^{1/2}} \exp\left(-\tfrac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

inverse (?)

determinant



if Σ not I, you get elipses

if Σ is I, similar to regular naive bayes, (however it does not have to be.)

$\mu \in \mathbb{R}^d$ (anything)

$\Sigma \in S^d_{++}$ symetric, positive definate

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \bar{x}_i$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (\bar{x}_i - \mu)(\bar{x}_i - \mu)^T + I$$

column vector

outer product of 2 vectors is a rank 1 matrix

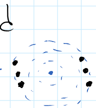
$\in S^d_{+}$ positive semi definate (some reason why no 2 $+$'s)

$\Sigma > 0$ strictly greater than zero

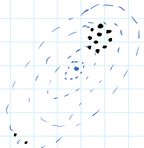
$$\int_{-\infty}^{\infty} p(x \mid \mu, \Sigma) = 1$$

Problems

1. $\hat{\Sigma}$ might not be valid
2. unimodal



3. Not robust (far away from mean)



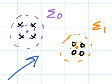
Gaussian Discriminant Analysis

$$p(\bar{x}_i, y_i) = p(\bar{x}_i \mid y_i)\, p(y_i)$$

$\bar{x}_i \in \mathbb{R}^d$

$y_i \quad \{1, 2, 3, 4\}$

$$(\bar{x}_i \mid y_i = c) \sim N(\mu_c, \Sigma_c)$$



$\mu_1 \quad \Sigma_1$
$\mu_2 \quad \Sigma_2$
$\mu_3 \quad \Sigma_3$
$\mu_4 \quad \Sigma_4$

know class, we are in some gaussian distribution

$$p(y_i \mid \bar{x}_i) = \alpha\, p(\bar{x}_i \mid y_i)$$

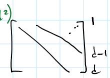
$p(\bar{x}_i)$

we don't care about $p(\bar{x}_i)$, does not depend on y, so we can already make a decision.

$$p(\bar{x}_i) = \sum_{c=1}^{C} \underbrace{p(\bar{x}_i \mid y_i = c)\, p(y_i = c)}_{p(\bar{x}_i,\, y_i = c)}$$

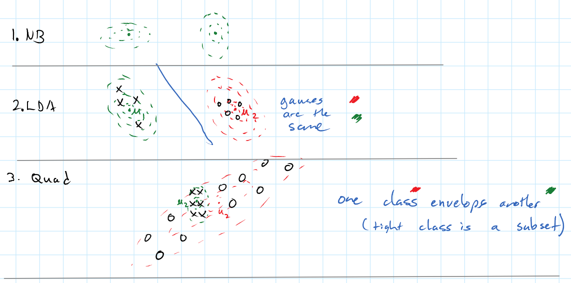
\# parameters

$$c + \underbrace{cd}_{\mu} + \overbrace{\frac{c(d+1)d}{2}}^{\Sigma} = O(cd^2)$$

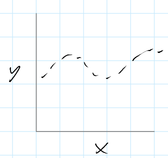


this is the sum of a matrix that is semetric : $\frac{d(d+1)}{2}$. there are c of the Σ terms.
there is also c of the μ terms. That gives the big O given, upper bounded by sigma ($d(d+1) = d^2$)

1. Naive Bayes assumption : Σ diagonal ~~isotropic,~~ axis aligned (circles of the same size)
2. Linear discriminant analysis: $\Sigma_c = \Sigma, \forall c$ each Σ similar to same common sigma
3. Quad Discr. anal. : general Σ_c



Gaussian Regression



both real

$$x = \begin{bmatrix} \cdots \end{bmatrix} \qquad y = \begin{bmatrix} \, \end{bmatrix}$$

"linear model"

"Linear regression" $y_i = w_1 x_i^1 + w_2 x_i^2 \ldots$ linear

$$y_i = w^T x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, 1)$$

$$y_i \mid \bar{x}_i, w \sim N(w^T \bar{x}_i, 1)$$

$$w_{MLE} = \underset{\bar{w} \in \mathbb{R}^d}{\arg\min} - \sum_{i=1}^{N} \log p(y_i \mid \bar{x}_i, \bar{w})$$

$$= \underset{\bar{w} \in \mathbb{R}^d}{\arg\min}\ \frac{1}{2} \sum_{i=1}^{n} (y_i - \bar{w}^T \bar{x}_i)^2$$ + constant (don't depend, doesn't change argmax)

"Least squares"

quadratic, multi dimensional

convex problem (so set δ/dx to 0 and solve)

$$X_w = \begin{bmatrix} \bar{w}^T \bar{x}_1 \\ \bar{w}^T \bar{x}_2 \\ \vdots \\ \bar{w}^T \bar{x}_n \end{bmatrix}$$

(inner product, sum over i of the square ?)

negative log likelyhood

targets

$$NLL(w) = \underset{\bar{w} \in \mathbb{R}^d}{\arg\min}\ \frac{1}{2} (Y - X_{\bar{w}})^T (Y - X_{\bar{w}})$$

$\Rightarrow \quad NLL = p(D \mid h)$

two of these terms, $2 \cdot \frac{1}{2}$

scalar $\quad NLL(w) = \frac{1}{2} Y^T Y - \bar{w}^T X^T y + \frac{1}{2} \bar{w}^T X^T X w$

with respect to

Partial δ/δ wrt variable)

d-dimensional vector $\quad \nabla NLL(w) = X^T y - X^T X \bar{w} = 0$ stationary

$$X^T X w = X^T y \qquad Aw = b \text{ linear}$$

d-dimensional matrix $\quad \nabla^2 NLL(w) = X^T X \quad$ convex (minimizer)