Maximum-Likelihood and Bayesian Parameter Estimation

- Sufficient Statistics
- Common Probability Distributions
- Problems of Dimensionality

Problems of Dimensionality

Problems involving 50 or 100 features (binary valued)

- Classification accuracy depends upon the dimensionality and the amount of training data
- Case of two classes multivariate normal with the same covariance

$$P(error\)=\frac{1}{\sqrt{2\pi}}\int\limits_{r/2}^{\infty}e^{\frac{-u^2}{2}}du$$
 where : $r^2=(\mu_1-\mu_2)^t\Sigma^{-1}(\mu_1-\mu_2)$
$$\lim_{r\to\infty}P(error\)=0$$

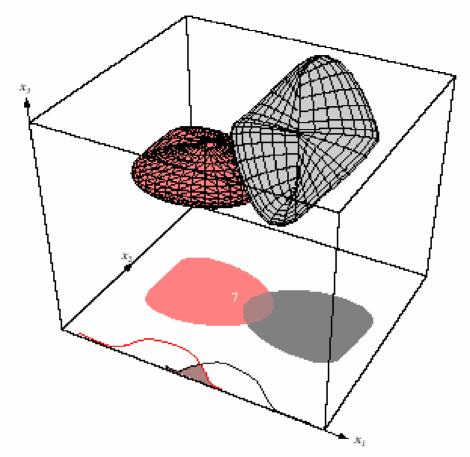
Error Rate and Dimensionality

If features are independent then:

$$\begin{split} \Sigma &= \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2) \\ r^2 &= \sum_{i=1}^{i=d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i}\right)^2 \end{split}$$

- Most useful features are the ones for which the difference between the means is large relative to the standard deviation
- It has frequently been observed <u>in practice</u> that, beyond a certain point, the inclusion of additional features leads to worse rather than better performance.

Decrease in error rate with features



Non-overlapping distributions in 3-dimensions where Bayes error vanishes. When projected to x_1 - x_2 sub-space or x_1 sub-space there is a greater overlap and hence greater Bayes error

7

Computational Complexity

Design methodology is affected by computational difficulty

"big oh" notation
$$f(x) = O(h(x))$$
 "big oh of $h(x)$ "

If: $\exists \mathbf{c}, \mathbf{x}_0 \ni |\mathbf{f}(\mathbf{x})| \le \mathbf{c}|\mathbf{h}(\mathbf{x})|$
 $\forall \mathbf{x} > \mathbf{x}_0$

(An upper bound on f(x) grows no worse than h(x) for sufficiently large x)

Example:
$$f(x) = 2+3x+4x^2$$

 $h(x) = x^2$ and c can be appropriately chosen
Therefore $f(x) = O(x^2)$

Big Theta Notation

"Big oh" is not unique

e.g.,
$$f(x) = O(x^2)$$
; $f(x) = O(x^3)$; $f(x) = O(x^4)$

Thus we introduce "big theta" notation

$$f(x) = \theta(h(x)) \text{ if: } \exists \mathbf{x}_0, \mathbf{c}_1, \mathbf{c}_2 \in \forall \mathbf{x} > \mathbf{x}_0 \\ 0 \le \mathbf{c}_1 \mathbf{h}(\mathbf{x}) \le \mathbf{f}(\mathbf{x}) \le \mathbf{c}_2 \mathbf{h}(\mathbf{x})$$

Therefore
$$f(x) = \theta(x^2)$$
 but $f(x) \neq \theta(x^3)$

Complexity of ML Estimation

Gaussian priors in d dimensions classifier with n training samples for each of c classes

For each category, we have to compute the discriminant

function
$$g(x) = -\frac{1}{2}(x - \widehat{\hat{\mu}}^{O(d.n)})^{O(n.d^2)} \underbrace{\sum_{i=1}^{O(n.d^2)} (x - \widehat{\mu}) - \frac{d}{2}\ln 2\pi}_{O(d^2.n)} + \underbrace{\ln P(\omega)}_{O(n)}$$

Total =
$$O(d^2 \cdot n)$$

Total for c classes = $O(cd^2 \cdot n) \cong O(d^2 \cdot n)$

Cost increase when d and n are large!

Overfitting

- Frequently, no of available samples usually inadequate, how to proceed?
- Solutions:
 - Reduce dimensionality: redesign feature extractor, or use a subset of features
 - Assume all c classes share same covariance matrix: pool available data
 - Assume statistical independence: all off-diagonal elements of covariance matrix are zero

10

• Classifier performs better than if we overfit the data, why?

Training Data from quadratic function plus Gaussian noise $f(x)=ax^2+bx+c+\varepsilon$ where $p(\varepsilon) \sim N(0,\sigma^2)$. Tenth order fits perfectly but second order performs better for new samples

Capacity of a Separating Plane

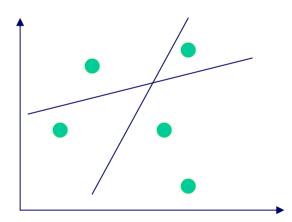
Overdetermined solution is significant for classification as it is for estimation

Task of partitioning a d-dimensional feature space by a hyperplane n samples in general position labelled either ω_1 or ω_2 . Of the 2^n possible dichotomies a fraction f(n,d) are linear dichotomies

$$f(n,d) = \begin{cases} \frac{1}{2} & n \le d+1 \\ \frac{2}{2^n} \sum_{i=0}^{d} {n-1 \choose i} & n > d+1 \end{cases}$$

Fraction of dichotomies of n points in d dimensions that are linear

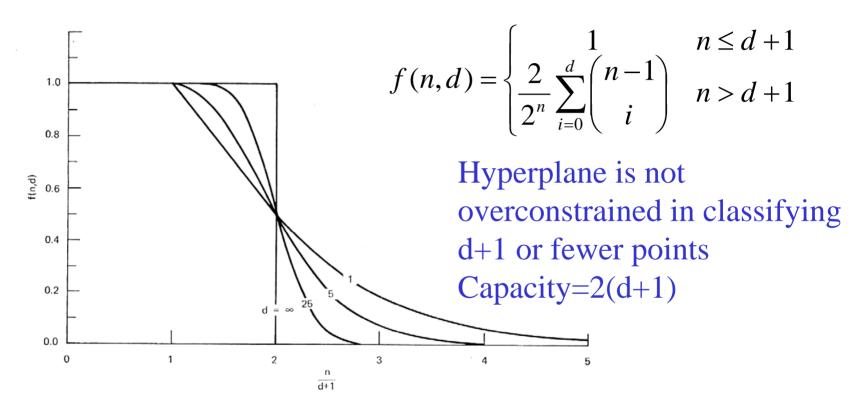
$$f(5,2) = 11/16$$



5 points can be labelled as either class in 32 ways

Capacity of a Separating Plane

Of the 2^n possible dichotomies a fraction f(n,d) are linear dichotomies



Fraction of dichotomies of n points in d dimensions that are linear