Organizational Comments

- We have a maillist --- please join ASAP.
- Reading for next week now on the schedule
- Vote for whether we insist on having presentations on-line.
- Tuesday we start regular format.
- Turnin keys are now set to cs645_NN where NN is the number of the day in the course that the response is for (schedule tells you the key).
  - To hand in a response file for day 04 (next Tuesday), sign onto the machine "lectura" and do:
    - turnin cs645_04 response.txt
- A few additional comments.
  - Most engagement with the technical material will be through self-study of the reading.
  - You need to engage with it seriously, but however it works for you.

Probabilistic Fitting

- Given the model, we have the probability of observing the data

$$p(D \mid \Theta) = \prod p(d_i \mid \Theta)$$

- But what we really want is the probability of the model (parameters) given the data!
- Bayes rule comes to the rescue!

- Bayes rule: $P(\Theta \mid D) = \dfrac{P(D \mid \Theta) P(\Theta)}{P(D)}$
- Proof $P(D, \Theta) = P(D \mid \Theta) P(\Theta) = P(\Theta \mid D) P(D)$

---

likelihood function for the parameters

prior probability (often taken to be uniform)

$$P(\Theta \mid D) = \frac{P(D \mid \Theta) P(\Theta)}{P(D)}$$

posterior probability

normalizer, often is not of interest

Common special case
$P(\Theta \mid D) \propto P(D \mid \Theta)$

Know the words in **red**

Probabilistic Fitting

- If we assume **uniform** prior, then we can find the posterior density for the parameters by:

$$p(\Theta \mid D) \propto p(D \mid \Theta)$$

- Now the objective is to find the parameters Θ such that this *likelihood* is maximum

Example One

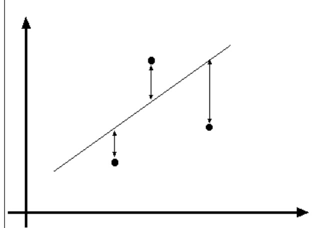
$D = ((x_1, y_1), (x_2, y_2), (x_3, y_3), \quad \ldots \quad (x_n, y_n))$ is dart locations.

$\Theta = (x_B, y_B)$ is the board center. Darts are distributed normally about the center, with covariance matrix σI.

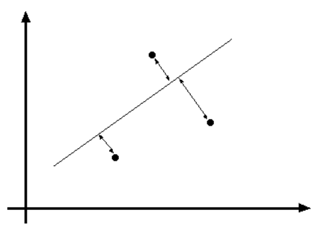
Assuming uniform prior, $p(\Theta \mid D) \propto p(D \mid \Theta)$

We want to find the value of Θ that maximizes $p(\Theta \mid D)$.

Example Two



Case one, no error in x's, "regression," like polynomials in Bishop chapter one.



Case two, error in both x and y.

In both cases, $p(D \mid \Theta) = \prod p(d_i \mid \Theta)$

Probabilistic fitting with independence and uniform prior

Finding the "best" model under simple circumstances

$\underset{\Phi}{\text{maximize}}\ \ p(\Theta \mid D)$ (one definition of best Θ)

$\underset{\Phi}{\text{maximize}}\ \ p(D \mid \Theta)$ (by Bayes rule, uniform prior)

$\underset{\Phi}{\text{minimize}}\ \ -\log(p(D \mid \Theta))$ (log is monotonic increasing)

$\underset{\Phi}{\text{minimize}}\ \ -\log\left(\prod p(d_i \mid \Theta)\right)$ (by independence)

$\underset{\Phi}{\text{minimize}}\ \ -\sum \log(p(d_i \mid \Theta))$ (high school math)

$$\underset{\Phi}{\text{minimize}}\ \ L(D \mid \Theta) = -\sum \log(p(d_i \mid \Theta)) \qquad \text{where} \quad \Theta = (x_B, y_B)$$

For darts

$$p(d_i \mid (x_B, y_B)) \propto \exp\left(-\left((x_i - x_B)^2 + (y_i - y_B)^2\right)\right)$$

$$-\log(p(d_i \mid (x_B, y_B))) = (x_i - x_B)^2 + (y_i - y_B)^2 + C$$

$$L(D \mid \Theta) = \sum_i (x_i - x_B)^2 + (y_i - y_B)^2$$

Minimize by taking derivatives with respect to x_B and y_B (separately is OK) and setting them to zero. For x_B we get:

$$0 = \sum_i (x_i - x_B) = \left(\sum_i x_i\right) - n$$

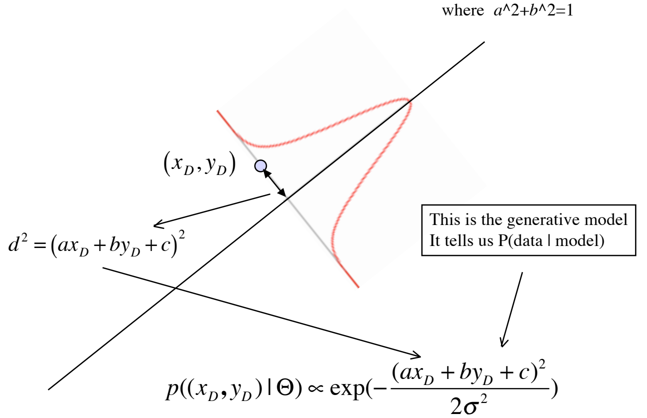
So, the ML solution is $x_B = \frac{1}{n}\sum_i x_i$

- For lines (case two), use ax+by+c=0 where a^2+b^2=1
- Algebraic fact: Distance squared from (x,y) to this line is (ax+by+c)^2
- **Generative model** for lines: Choose point on line, and then, with probability proportional to p(d), **normally distributed** (Gaussian), go a distance d from the line.
- Now the probability of an observed (x,y) is given by

$$p((x,y) \mid \Theta) \propto \exp\left(-\frac{(ax+by+c)^2}{2\sigma^2}\right)$$

Lines

Convenient formula for line
ax+by+c=0
where *a*^2+*b*^2=1



We have the probability density of the observed (x,y) given by

$$p((x,y) \mid \Theta) \propto \exp\left(-\frac{(ax+by+c)^2}{2\sigma^2}\right)$$

The negative log is

$$\frac{(ax+by+c)^2}{2\sigma^2}$$

And the negative log likelihood of multiple observations is

$$\frac{1}{2\sigma^2}\sum_i (ax_i + by_i + c)^2$$

From the previous slide, we had that the negative log likelihood of multiple observations is given by

$$\frac{1}{2\sigma^2}\sum_i (ax_i + by_i + c)^2 \qquad \text{(where } a^2 + b^2 = 1\text{)}$$

This is known as homogeneous least squares which can be solved using eigenvalues.