Probability Concepts

- We will make use of the following concepts.
  - Basic probability in discrete spaces, events
  - Joint probability
  - Conditional probability
  - Independence (and conditional independence)
  - Marginal probability (marginalization)
  - Probability in continuous spaces (probability density functions)

- To learn/review them see supplementary chapter in the book posted on the web site or you favorite web text or video resource

Probabilistic Fitting

- Generative probabilistic model
  - Tells a story about how stochastic data comes to be
  - Darts fall around the center of the board, but where exactly?
  - Consider a model with parameters, $\boldsymbol{\theta}$
  - Consider an observation, x_i
  - We denote the probability of seeing x_i under the model by:

$$p(x_i \mid \Theta)$$

Read "given" or "conditioned on"
Restricts to the case of $\boldsymbol{\theta}$

Defined by $P(A \mid B) = \frac{P(A,B)}{P(B)}$

Probabilistic Fitting

- Multiple observations
  - Suppose we have multiple observations, in a vector $\mathbf{x}$
  - What is the probability of $\mathbf{x}$?
- If observations are independent then probability is the product of the individual observations
  - Essentially a definition, but it is consistent with intuition
  - The observations are conditionally independent **given** the model
- So, the probability of $\mathbf{x}$ is then:

$$p(\mathbf{x} \mid \Theta) = \prod p(x_i \mid \Theta)$$

Probabilistic Fitting

- So, given the model, we have the probability of observing the data

$$p(\mathbf{x} \mid \Theta) = \prod p(x_i \mid \Theta)$$

- But what we really want is the probability of the model (parameters) given the data!

- Bayes rule comes to the rescue!

Bayes Rule

- Bayes rule: $P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$
- Proof $P(A,B) = P(B \mid A)P(A) = P(A \mid B)P(B)$
- With our notation: $P(\Theta \mid \mathbf{x}) = \dfrac{P(\mathbf{x} \mid \Theta)P(\Theta)}{P(\mathbf{x})}$

likelihood function
for the parameters

prior probability (often
taken to be uniform)

$$P(\Theta \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \Theta)P(\Theta)}{P(\mathbf{x})}$$

posterior probability

normalizer, often is
not of interest

Common special case
$P(\Theta \mid \mathbf{x}) \propto P(\mathbf{x} \mid \Theta)$

Know the words in **red**

Probabilistic Fitting

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

$$p(\Theta \mid \mathbf{x}) \propto p(\mathbf{x} \mid \Theta)$$

- Now the objective is to find the parameters Θ such that this *likelihood* is maximum
- Note--this is the same as finding the parameters which minimize the **negative log likelihood**

Probabilistic fitting with independence and uniform prior

Finding the "best" model under simple circumstances

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

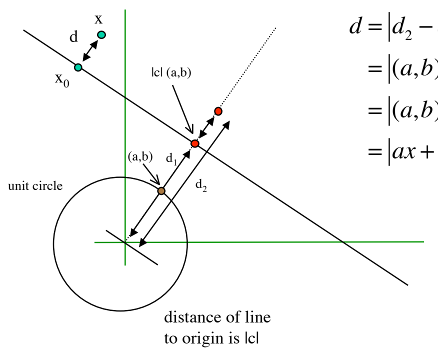
$\underset{\Phi}{\text{maximize}}\ \ p(\mathbf{x} \mid \Theta)$ (by Bayes rule, uniform prior)

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

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

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

- Back to lines: ax+by+c=0 where a^2+b^2=1
- Distance squared from (x,y) to this line is (ax+by+c)^2



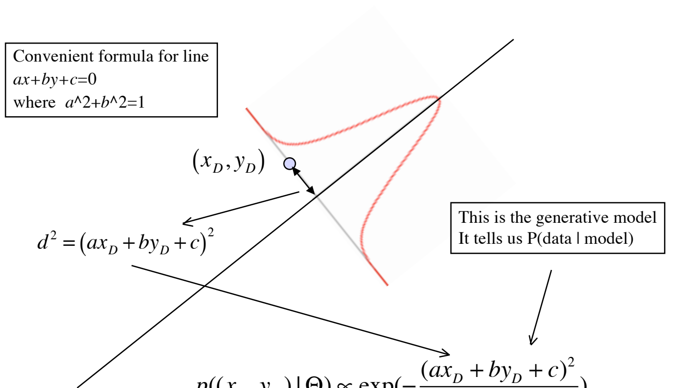
$$\begin{aligned} d &= \left|d_2 - d_1\right| \\ &= \left|(a,b)\bullet\mathbf{x} - (a,b)\bullet\mathbf{x_0}\right| \\ &= \left|(a,b)\bullet\mathbf{x} + c\right| \\ &= \left|ax + by + c\right| \end{aligned}$$

- **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

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

Lines



$$p((x_D, y_D)\,|\,\Theta) \propto \exp\left(-\frac{(ax_D + by_D + c)^2}{2\sigma^2}\right)$$

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

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

The negative log is

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

And the negative log likelihood of multiple observations is

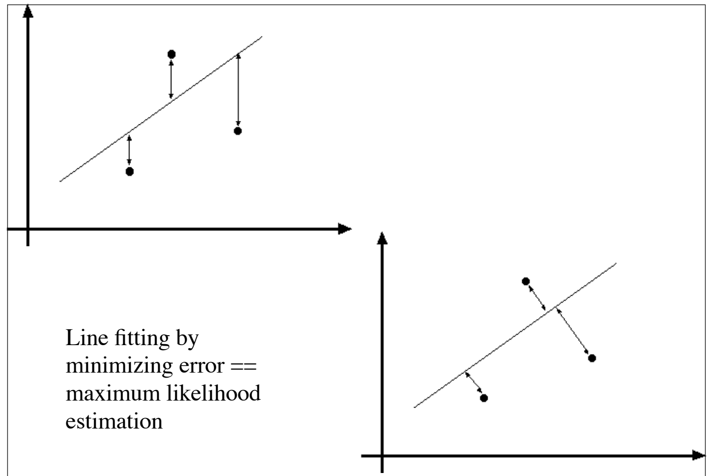
$$\frac{1}{2\sigma^2}\sum_i (ax_i + by_i + c)^2$$

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 should be recognizable as homogeneous least squares

Thus we have shown that least squares is maximum likelihood estimation under normality (Gaussian) error statistics!



Line fitting by minimizing error == maximum likelihood estimation

Fitting curves other than lines

- In principle, an easy generalization
  - For Gaussian error statistics, Euclidean distance is a good measure
  - The probability of obtaining a point, given a curve, is given by a negative exponential of distance squared
- In practice, this can be hard
  - It can be difficult to compute the distance between a point and a curve
  - Circles, ellipses, and a few others are not too hard
  - Otherwise, craft an approximation
  - §15.3 has more