

CS 696i, Spring 2005, part III

Computer Vision

(Making Machines See)

and other miscellaneous stuff

Kobus Barnard

Computer Science, University of Arizona

Statistical Methods / Machine Learning

Why probabilistic methods?

Statistical Methods / Machine Learning

Why probabilistic methods?

Need to generalize (predict) to beat the competition

Statistical Methods / Machine Learning

Why probabilistic methods?

Need to generalize (predict) to beat the competition

The competition “measures” how you do as a function of the statistics of the world (test data).

This suggests statistical methodology to get the “best” answer.

Statistical Methods / Machine Learning

Why models?

Statistical Methods / Machine Learning

Why models?

Need to simplify---a serviceable model that is simpler than the data gives a competitive advantage.

Statistical Methods / Machine Learning

Why models?

Need to simplify (a serviceable model that is simpler than the data gives a competitive advantage because the possible data is huge).

What makes a serviceable model?

Statistical Methods / Machine Learning

Why models?

Need to simplify (a serviceable model that is simpler than the data gives a competitive advantage because the possible data is huge).

What makes a serviceable model?

It captures that which is useful for prediction. The stuff which is not modeled is implicitly irrelevant (noise or variation that does not serve your purpose).

Terminology

Generative statistical model

Mathematically generates data samples

Training/learning

Figuring out the model from example (training) data

Inference

Using the model on new data to predict

Terminology

Held out data

Data hidden from the training process used to test generalization

Cross validation

Measuring performance on held out data, often in a specified pattern, e.g., leave one out, or K-fold.

IID

Identically, independently distributed.

The Bayesian Way

Features of the Bayesian Way

- Uses Bayes rule

- The model and data are treated statistically alike

- Prior probabilities

- Postpone making decisions (behave as though you have access to probability distributions)

- Risk functions

Bayes Rule

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

Bayes Rule

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model}) P(\text{model})}{P(\text{data})}$$

Bayes Rule

By definition, $P(a|b) = P(a,b) / P(b)$

Bayes Rule

By definition, $P(a|b) = P(a,b) / P(b)$

so $P(a, b) = P(a | b) * P(b)$

Bayes Rule

By definition, $P(a|b) = P(a,b) / P(b)$

so $P(a, b) = P(a | b) * P(b)$

and $P(a, b) = P(b | a) * P(a)$

Bayes Rule

By definition, $P(a|b) = P(a,b) / P(b)$

so $P(a, b) = P(a | b) * P(b)$

and $P(a, b) = P(b | a) * P(a)$

so $P(a | b) * P(b) = P(b | a) * P(a)$

Bayes Rule

By definition, $P(a|b) = P(a,b) / P(b)$

so $P(a, b) = P(a | b) * P(b)$

and $P(a, b) = P(b | a) * P(a)$

so $P(a | b) * P(b) = P(b | a) * P(a)$

and
$$P(a | b) = \frac{P(b | a)P(a)}{P(b)}$$

Bayes Rule

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model}) P(\text{model})}{P(\text{data})}$$

Bayes Rule

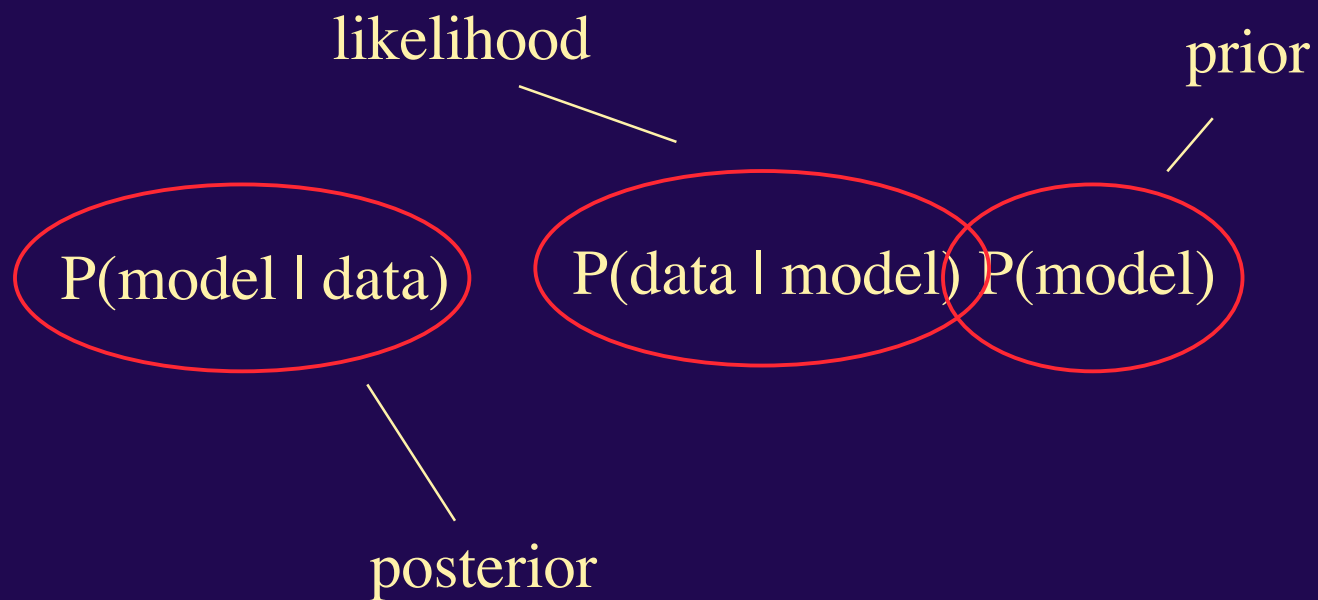
Diagram illustrating Bayes Rule:

$$\text{P(model | data)} = \frac{\text{P(data | model) P(model)}}{\text{P(data)}}$$

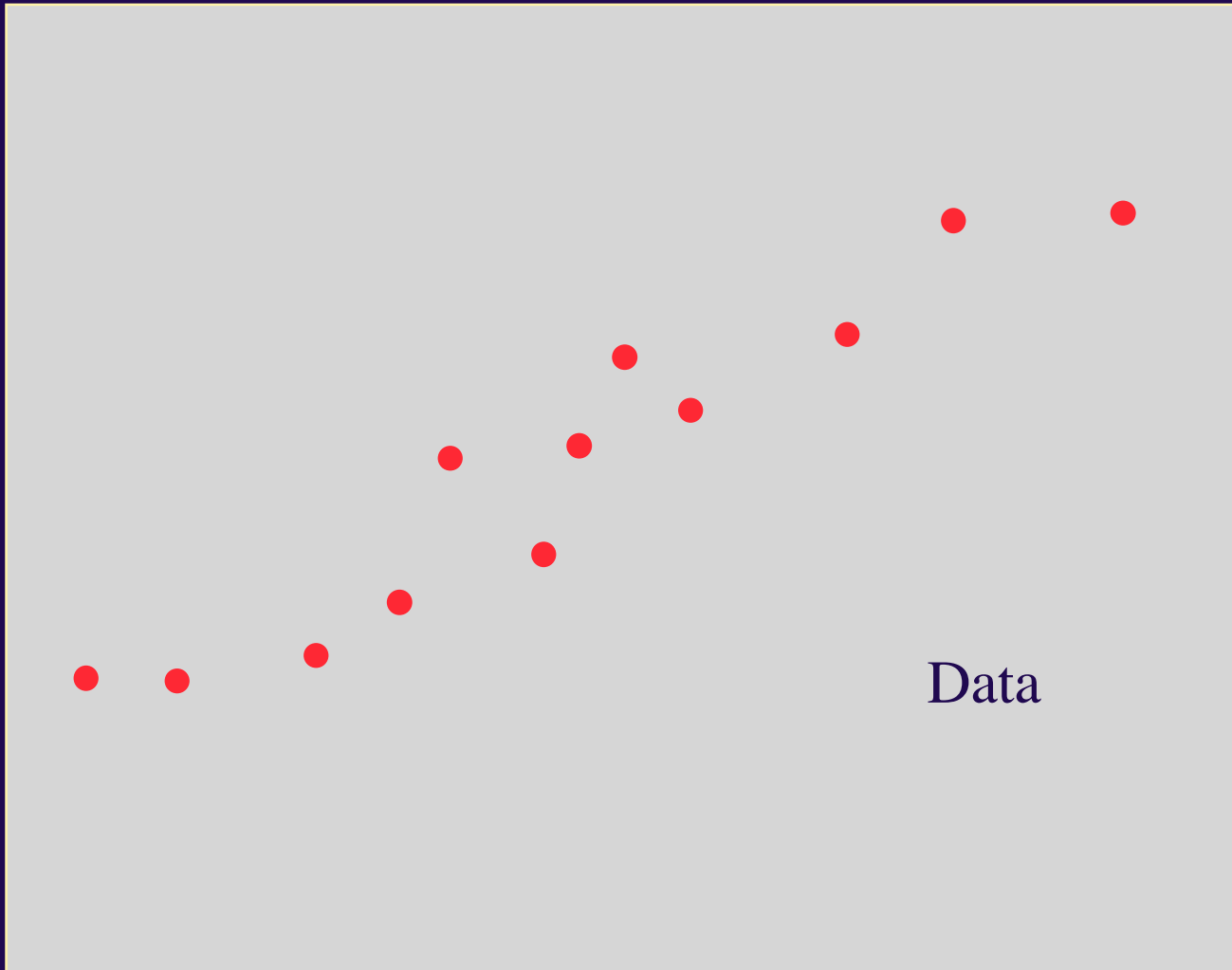
The diagram includes the following labels and annotations:

- likelihood**: Points to P(data | model)
- prior**: Points to P(model)
- posterior**: Points to P(model | data)
- Red ovals highlight the terms P(model | data) , P(data | model) , and P(model) .

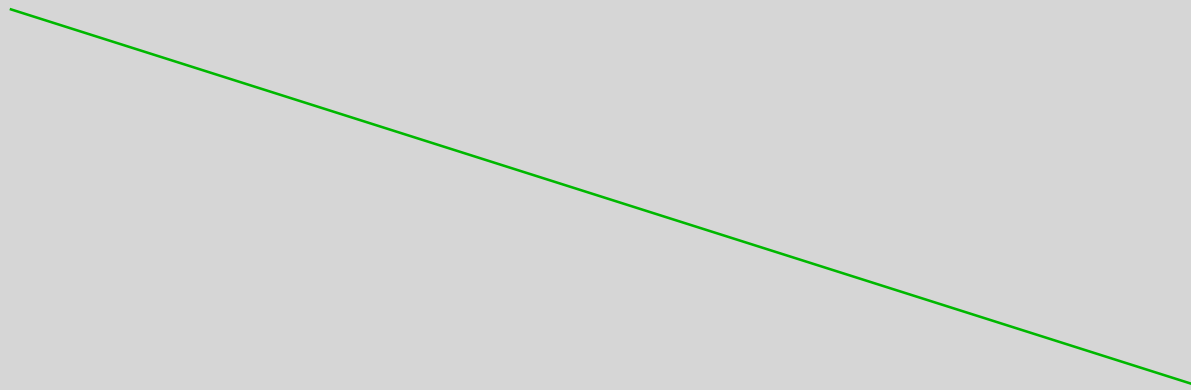
Bayes Rule



Example: Line Fitting



Example: Line Fitting



Model instance
(with particular parameters,
e.g. slope and intercept)

Generative Model

Represent lines by:

$$au+bv+c=0$$

$$\text{where } a^2+b^2=1$$

Algebraic fact:

Distance squared from a point (x,y) to this line is
 $(ax+by+c)^2$

Generative Model

Represent lines by:

$$a \bullet x + b \bullet y + c = 0$$

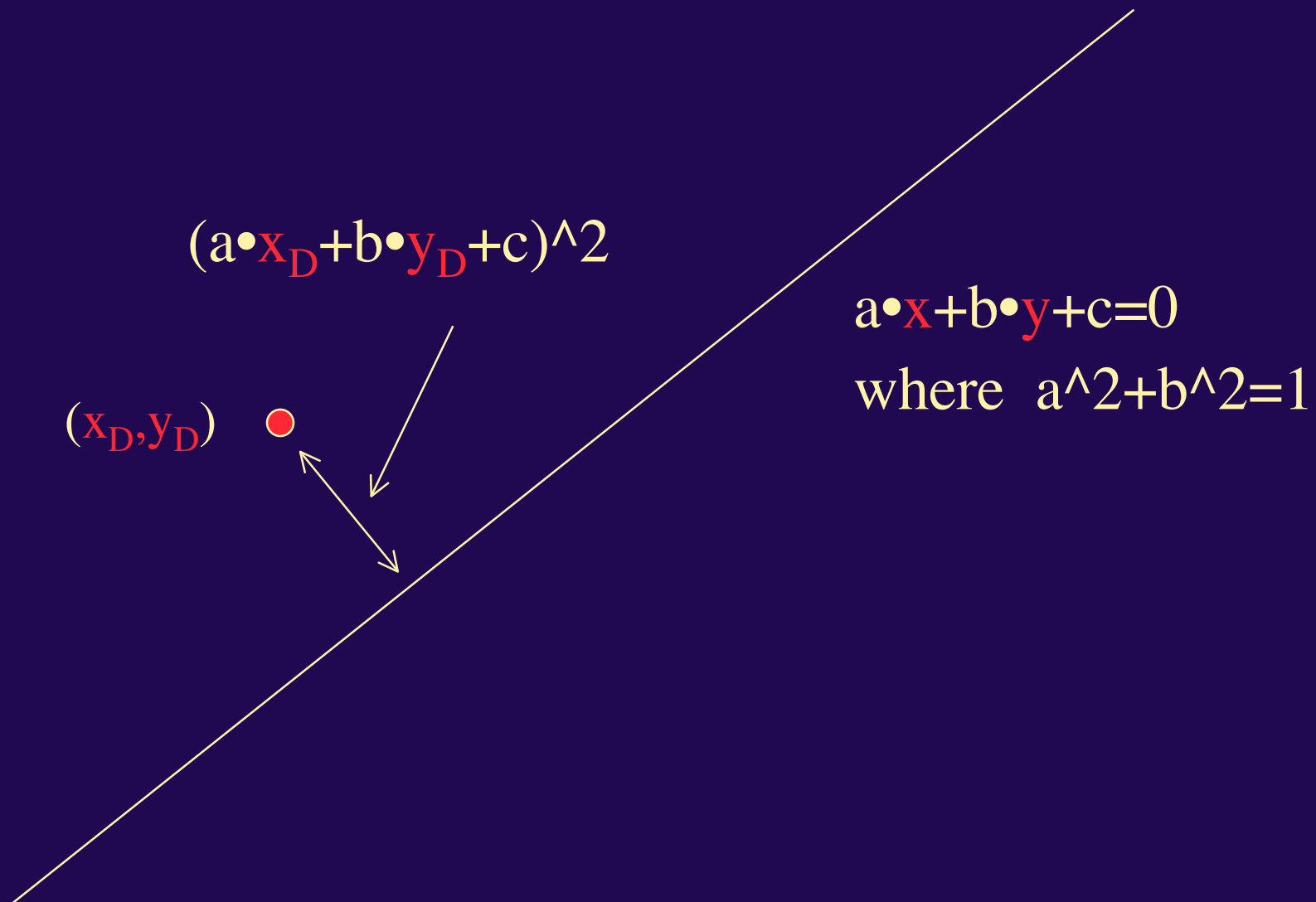
$$\text{where } a^2 + b^2 = 1$$

Algebraic fact:

Distance squared from a point (x_D, y_D) to this line is

$$(a \bullet x_D + b \bullet y_D + c)^2$$

Generative Model



Generative Model

To generate the data

1. Choose (according to prior distribution), (a,b)
Equivalent to choosing the slope and intercept.
2. Now generate data points with probability so that the closer they are to the line, the more likely.

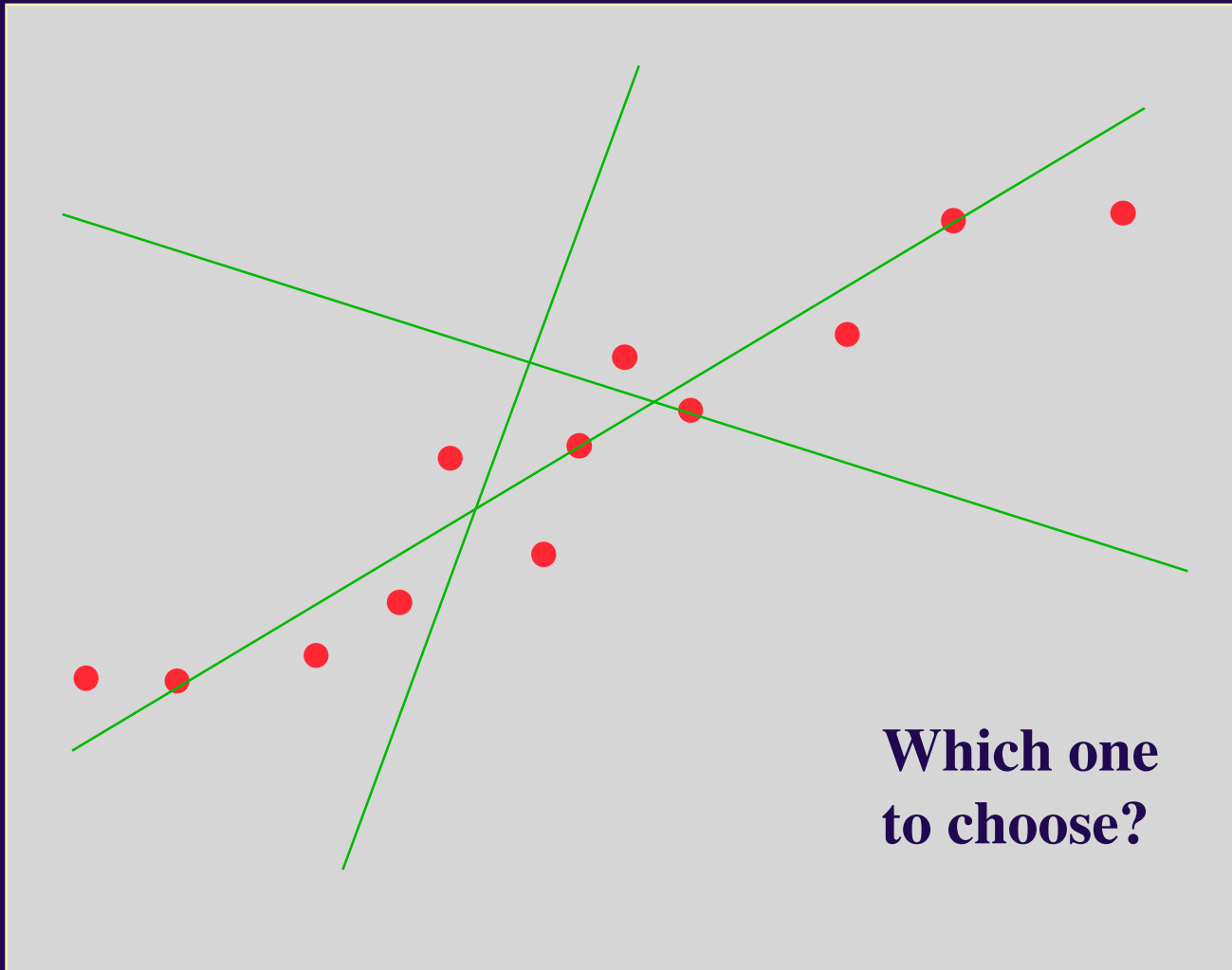
Generative Model

To generate the data


1. Choose (according to prior distribution), (a,b)
Equivalent to choosing the slope and intercept.
2. Now generate data points with probability so that the closer they are to the line, the more likely.

In particular, let's say that data points are Normally distributed as a function of the distance to the line.

Example: Line Fitting



Now the probability of an observed (x,y) is given by

$$P((x, y) | \square) = \exp\left(-\frac{(ax + by + c)^2}{2\square^2}\right)$$


Common notation for model parameters, here (a,b).

Now the probability of an observed (x,y) is given by

$$P((x,y) | \beta) = \exp\left(-\frac{(ax + by + c)^2}{2\beta^2}\right)$$

The probability of all the observations (i.i.d) is

$$\prod_i P((x_i, y_i) | \beta)$$

Substituting

$$\prod_i \exp\left(-\frac{(ax_i + by_i + c)^2}{2\beta^2}\right)$$

Turn the product into a sum by taking logs:

$$\log(P(data|\theta)) = \log\left(\prod_i P((x_i, y_i)|\theta)\right) = \sum_i \log(P((x, y)|\theta))$$

Recall that:

$$P((x, y)|\theta) = \exp\left(-\frac{(ax + by + c)^2}{2\theta^2}\right)$$

So, ignoring constants, the negative of the data log likelihood is:

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

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

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

If we want the likeliest line, we maximize the log likelihood which is the same as minimizing the above.

We could solve this by considering derivatives, but this is recognizable as homogeneous least squares

Thus we have shown that least squares is the maximum likelihood estimate of the line under normality (Gaussian) error statistics!

Important Notes

While least squares runs out of steam, the general approach carries on.

We are able to get a maximum likelihood estimate, but in fact, we had access to the entire probability distribution of the model, given the data.

Using this distribution, we can compute:

- Compute other estimates of “best”

- Compute estimates using a risk or cost function

- Use the distribution itself for further inference

More Model Examples

Clustering:

Select a cluster, with probability $P(c)$

Generate data with cluster dependent probability
 $P(\text{obs} | c)$

Outliers process:

Choose regular model or outlier model

Generate differently depending on case

More Model Examples

Recognition:

Select object with $P(o)$

Based on the object, generate parts, p , with probability
 $P(p \mid o)$

For each part, p , generate features, f :
 $P(f \mid p)$

Inference: Given the model parameters for each object and observed features, compute the likelihood of each object.

Inference

Having learned a model, we can compute the likelihood of new data

Typical processes:

- Compute the joint probability of input and output

- Compute the likelihood of each class (recognition)

Generative vs Discriminative models

