ECSE-626
Statistical Computer Vision

Probabilistic Inference

What is a probability?

- There is an ongoing debate about the definition of a probability.
- Basically 2 camps: frequentists vs. Bayesians.

Frequentist view

- Frequentists dominated field of statistics for most of 20th century.
- Frequentists describe probabilities as: *frequencies of outcomes in random experiments*.
- E.g. Probability that a coin comes up heads – average fraction of heads if you perform a long sequence of coin flips.

Bayesian View

- Probabilities describe the *degree of belief* in propositions: e.g. the probability that Mr. S was a murderer given the evidence.
- Probabilities can describe inferences about the world.
- Aren't really worried about what is the "true" state of the world since we will never really know if we are right.
- All we can state is the probability of a hypothesis given the evidence.

Bayesian View

- Also, known as *subjective* interpretation of probability, since probabilities depend on assumptions.
- Frequentists see this as a problem.

Bayesian View

- Advocates of Bayesian approach to modelling and pattern recognition don’t see subjectivity as a problem.
- In their view, you can’t perform inference without making assumptions. The question is whether you make your assumptions explicit or if you “sweep them under the rug”.

Cox's Axioms

- Not all degrees of belief that add up to 1 are true probabilities.
- If a set of beliefs satisfy Cox's axioms, a set of simple consistency rules, then they can be called *probabilities*.
- Probabilities follow certain mathematical protocols.

Cox's Axioms

- Denote degree of belief in a proposition x, B(x).

  Axiom 1: Degrees of belief can be ordered: If $B(x)$ is 'greater' than $B(y)$, and $B(y)$ is greater than $B(z)$, then $B(x)$ is greater than $B(z)$.

Cox's Axioms

Axiom 2: Degree of belief in a proposition x and it negation $\bar{x}$ are related. There is a function f such that:

$$B(x) = f\,[B(\bar{x})].$$

Cox's Axioms

Axiom 3: The degree of belief in a conjunction of propositions x, y (x AND y) is related to the degree of belief in the conditional proposition $x|y$ and the degree of belief in the proposition y. There is a function g such that:

$$B(x,y) = g[B(x|y), B(y)]$$

Review of Some Rules

- Ensemble: X is a triple (x, Ax, Px) where the outcome x is that value of a random variable, which takes on one of a set of possible values, $Ax = \{a_1, a_2, \ldots, a_i, \ldots, a_I\}$, having probabilities $Px = \{p_1, p_2, \ldots, p_I\}$, with $P(x=a_i) = p_i$, $p_i \geq 0$ and

$$\sum_{a_i \in Ax} P(x = a_i) = 1$$

Review of Some Rules

- Marginal probability: We can obtain the marginal probability $P(x)$ from the joint probability $P(x,y)$ by summation:

$$P(x = a_i) \equiv \sum_{y \in A_X} P(x = a_i, y)$$

or

$$P(x) \equiv \sum_{y \in A_X} P(x, y)$$

Review of Some Rules

- Conditional probability:

$$P(x = a_i \mid y = b_j) \equiv \frac{P(x = a_i, y = b_j)}{P(y = b_j)}$$

if

$$p(y = b_j) \neq 0$$

Review of Some Rules

- Product rule:

$$P(x, y) = P(x \mid y)P(y)$$

$$= P(y \mid x)P(x)$$

Review of Some Rules

- Sum rule:

$$P(x) = \sum_{y} P(x, y)$$

$$= \sum_{y} P(x \mid y) P(y).$$

Bayes' Theorem

Thomas Bayes
(1702 – 1761)

http://en.wikipedia.org/wiki/Thomas_Bayes

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)}$$

$$= \frac{P(x \mid y)P(y)}{\sum_{y'} P(x \mid y')P(y')}$$

Review of Some Rules

- Independence: Two random variables X and Y are *statistically independent* iff

$$P(x, y) = P(x)P(y).$$

Review of Some Rules

- It is important NOT to be sloppy regarding the definitions and notations related to probability theory.
- Probability distribution D(x) or cumulative density function (CDF) describes the probability that a random variable X takes on a value less than or equal to a number x, $P(X \le x)$.
- For a continuous distribution:

$$D(x) = P(X \le x) = \int_{-\infty}^{x} p(x)dx$$

Review of Some Rules

- Probability density function (PDF), $p(x)$, shows how the density of possible observations is distributed.
- It is the derivative of the distribution function of a random variable:

$$D'(x) = p(x)$$

Review of Some Rules

- Note that:

$$P(x \in B) = \int_B p(x)dx$$

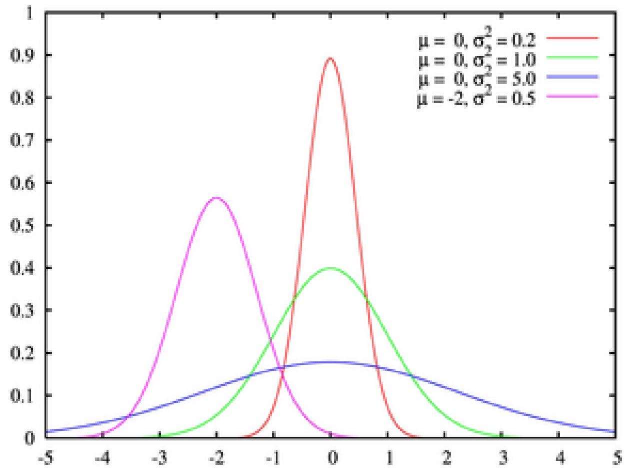
$$P(-\infty < x < \infty) = \int_{-\infty}^{\infty} p(x)dx = 1$$

$$P(x = a) = \int_a^a p(x)dx = 0.$$

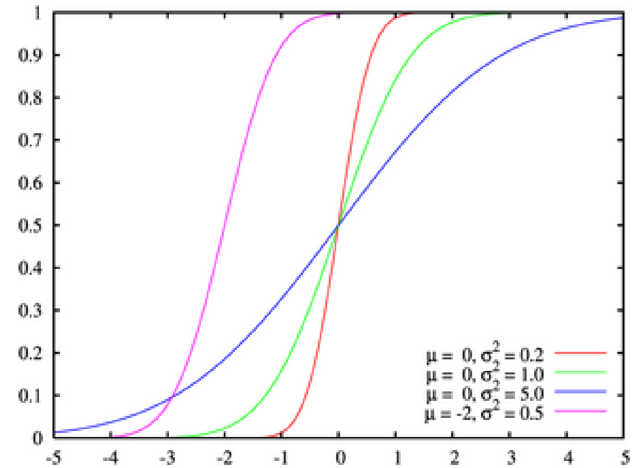
Review of Some Rules

Normal distribution

Probability density function



Probability distribution or CDF



http://en.wikipedia.org/wiki/Normal_distribution

Forward probability vs. inverse probability

- Forward probability problems involve *generative model*: describe the process that is assumed to give rise to some data.
- Task is to compute the probability distribution or expectation of some quantity that depends on the data.
- Want to predict behaviour of the data given the state.

Forward Probability

- Let us define the following:

  θ denotes the unknown parameters we are trying to infer,

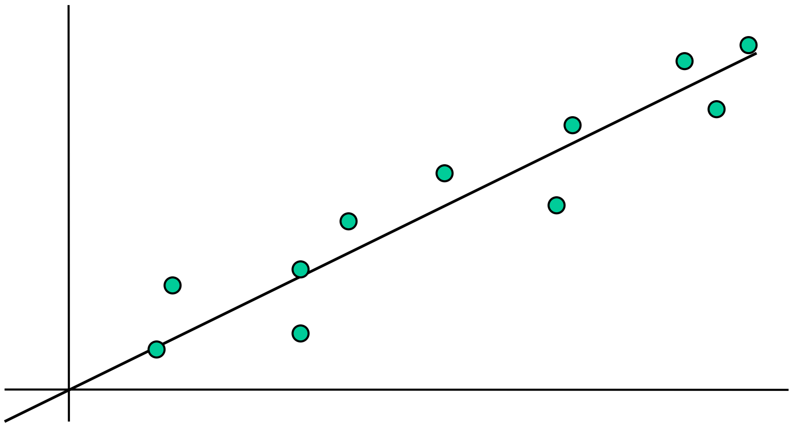
  D denote the data measurements.

- We are trying to estimate the probability density function:

$$p(D \mid \theta)$$

Forward Probability

Example: Estimate the distribution of the data points, given the parameters of a straight line model.



D: data points
Θ: parameters of line equation

Estimate the probability density function:

$$p(D \mid \theta)$$

Inverse Probability

- *Inverse probability* problems involve a generative model of a process as well:
  - Instead of computing the probability distribution of some quantity *produced by* the process, compute the conditional probability of one or more of the *unobserved variables* in the process, *given* the observed variables.
  - E.g. Infer the parameters of the line equation given a set of data measurements.

$$p(\theta \mid D)$$

Bayesian Inference

- This can be solved by application of Bayes' Law:

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}$$

$p(D \mid \theta)$	Likelihood function
$p(\theta)$	Prior density function on θ
$p(D)$	Marginal density function on D
$p(\theta \mid D)$	Posterior density function of θ given D.

Likelihood

- Likelihood function, $p(D \mid \theta)$, not always a probability distribution – refer to the *likelihood of the parameters* (not the data).
- For fixed parameters: describes the probability of the data, D, given the parameters, θ.
- Defines the forward probability.

Likelihood

- In many inference tasks, defines the *physical theory*: describes the relationship between the physical measurements as acquired by a sensor and the parameters to be estimated.
- In computer vision (and other tasks), the forward probability distribution is often computed during a *learning* or *training* phase.

Prior

- The prior distribution describes the state of information prior to any data arriving.
- In our example, $p(\theta)$ describes the marginal density function on the parameters, prior to data, D, being acquired.
- Also referred to as the *subjective prior* – since it explicitly embeds assumptions about the state of information prior to data arriving.

Normalization

- The marginal density function, $p(D)$, is sometimes called the normalization constant, since it will always be the same, regardless of the alternate set of parameters estimated.

Posterior

- The conditional probability density function: $p(\theta \mid D)$ is called the posterior probability density function of the parameters given the data.
- It describes how the parameters estimates change after the data arrive.
- The inverse probability problem involves estimating this distribution from measurements.

Bayesian Inference

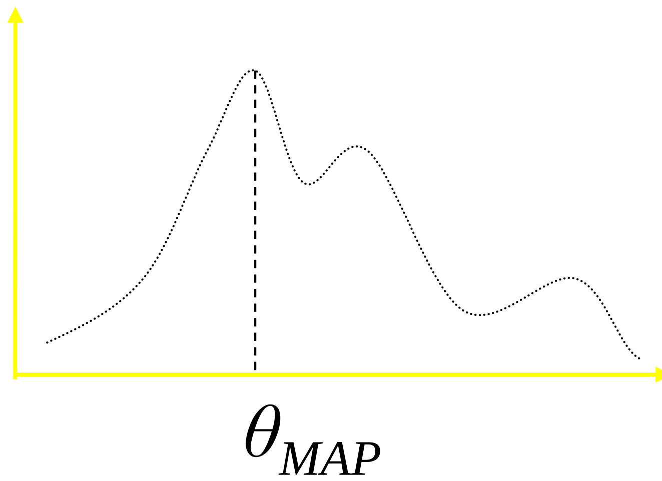
- The beauty of the Bayesian approach is that all sources of uncertainty are made explicit, in the form of probability density function. There are no hidden assumptions.
- It provides a recipe for inverse problems that can be used in a wide variety of applications.
- Changes can be made to the form of the distributions without changing the entire framework.

Bayesian Inference

- The result is the form of a probability distribution rather than a single solution.
- The posterior probability distribution describes the degree of confidence in various solutions. It is up to higher level processes to then determine what to do with the distribution.

Bayesian Inference

$$p(\theta \mid D)$$



• Often you want to choose a single solution – not a problem! Can choose the *Maximum A Posteriori* solution (MAP) – the one that the system has the highest confidence in.

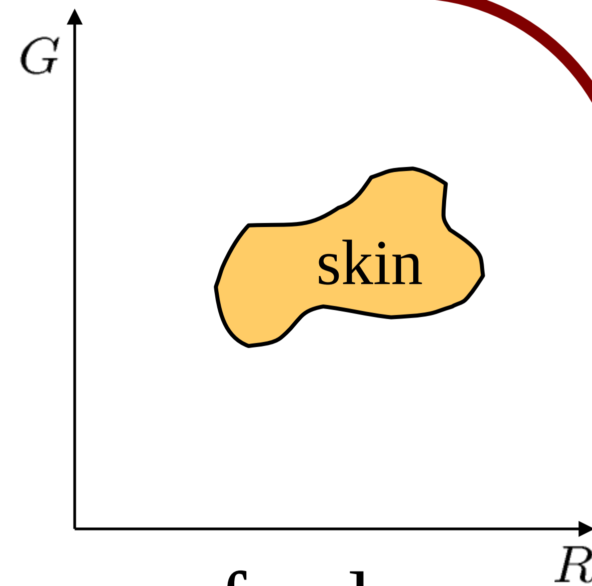
Bayesian Inference in vision

- In this course, we will focus on posing problems in computer vision as probabilistic inference problems.
- Let's look at an example..

Let’s start with skin detection

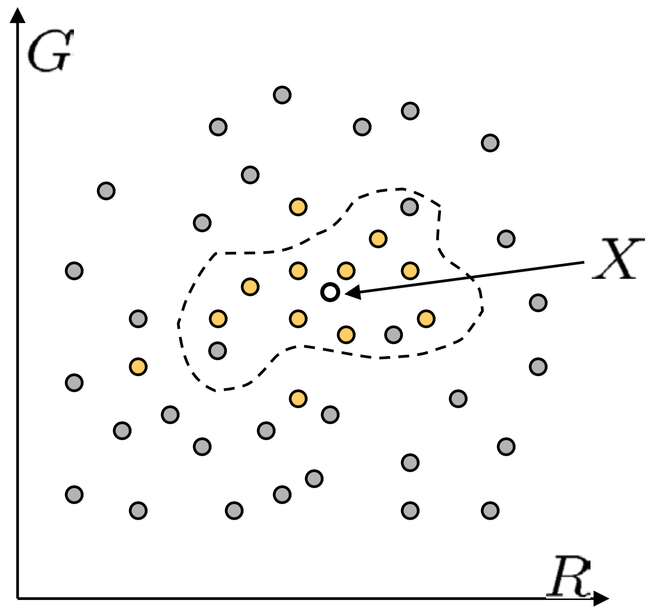
popscan.blogspot.com www.wordy.photos

Skin Detection



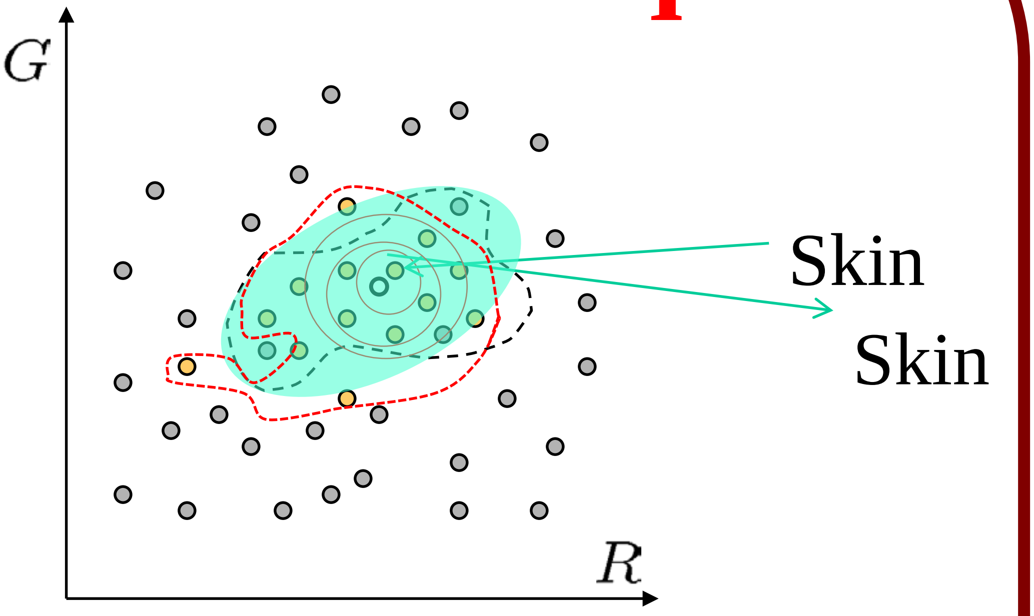
- Skin pixels have a distinctive range of colors
  - Corresponds to region(s) in RGB color space
    - for visualization, only R and G components are shown above
- Skin classifier
  - A pixel X = (R,G,B) is skin if it is in the skin region
- But how to find this region?

Skin Detection



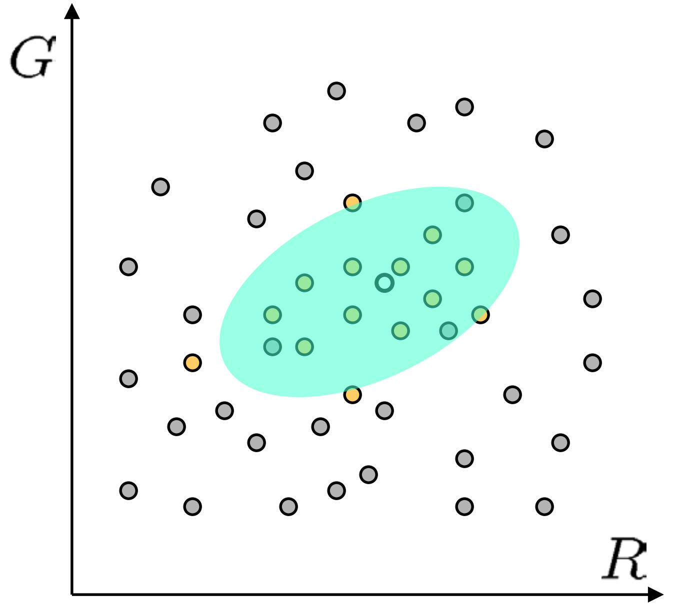
- **Learn** the skin region from examples
  - Manually label pixels in one or more “training images” as skin or not skin
  - Plot the training data in RGB space
    - skin pixels shown in yellow, non-skin pixels shown in black
    - some skin pixels may be outside the region, non-skin pixels inside. Why?

Skin classification techniques



- Skin classifier
  - Given X = (R,G,B): how to determine if it is skin or not?
  - Nearest neighbor
    - find labeled pixel closest to X
    - choose the label for that pixel
  - Data modeling
    - Model the *distribution* that generates the data (Generative)
    - Model the *boundary* (Discriminative)

Skin classification techniques

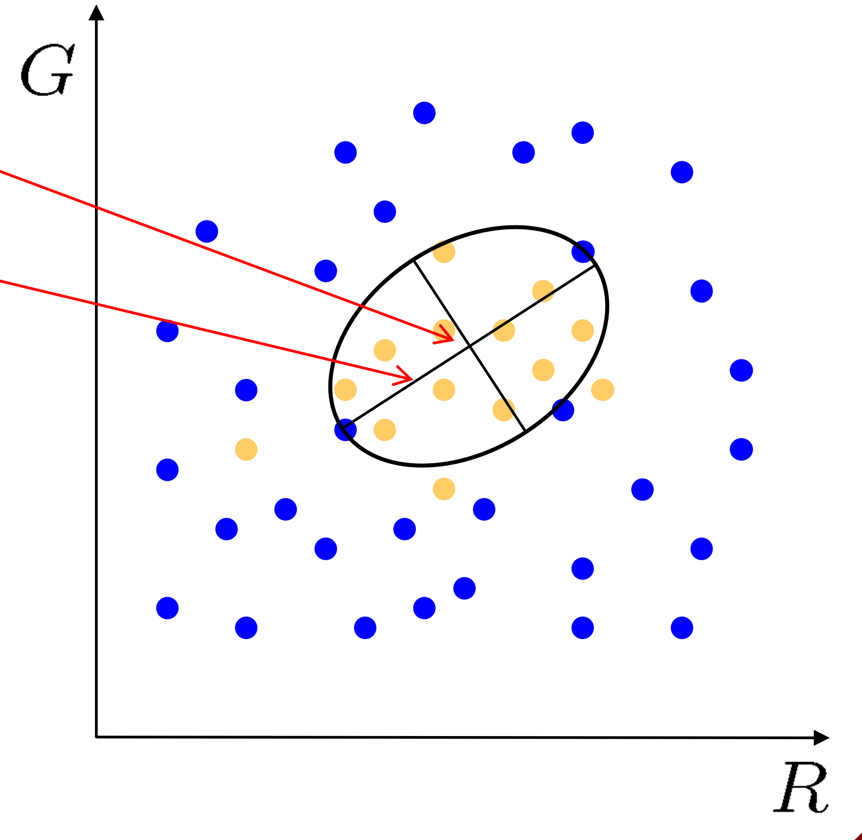


- We can fit a probability distribution to model the skin samples
  - E.g. Gaussian

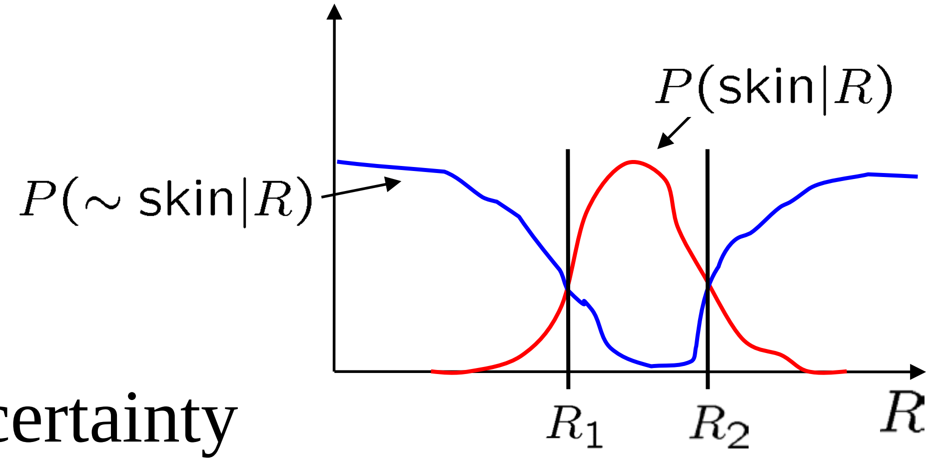
Fitting a Gaussian to Skin samples

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

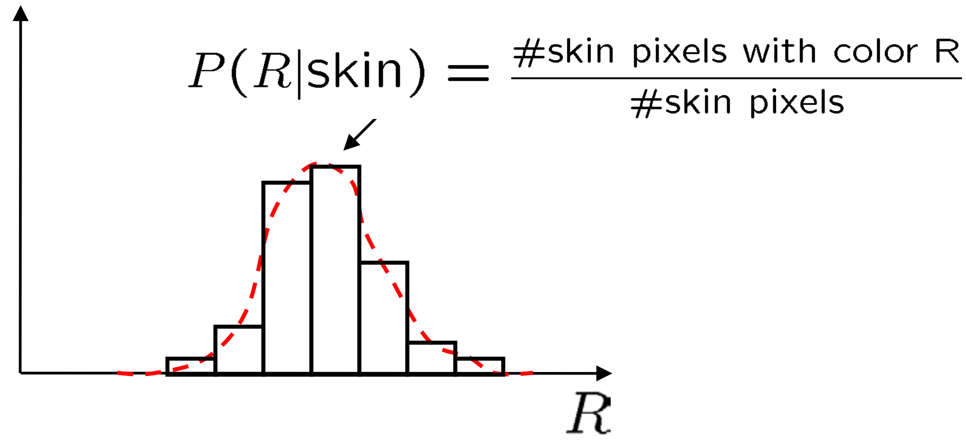


Probabilistic skin classification



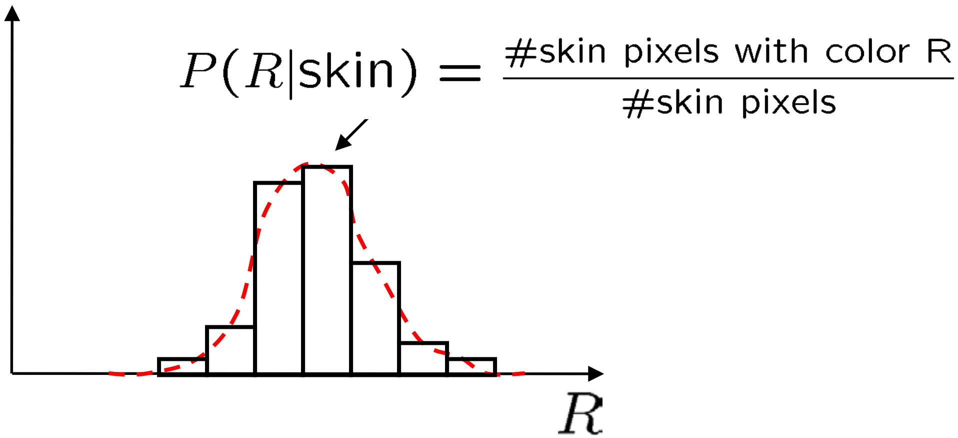
- Now we can model uncertainty
  - Each pixel has a probability of being skin or not skin
  - $P(\sim \text{skin}|R) = 1 - P(\text{skin}|R)$
- Skin classifier
  - Given X = (R,G,B): how to determine if it is skin or not?
  - Choose interpretation of highest probability
  - set X to be a skin pixel if and only if $R_1 < X \leq R_2$
- Where do we get $P(\text{skin}|R)$ and $P(\sim \text{skin}|R)$

Learning conditional PDF's



- We can calculate p(R | $skin$) from a set of training images
  - Approach: fit parametric PDF functions
    - common choice is Gaussian

Learning conditional PDF's



- We can calculate p(R | $skin$) from a set of training images
- We want p($skin$ | R) not p(R | $skin$)
- How can we get it?

Bayesian estimation

what we measure
(**likelihood**)

domain knowledge
(**prior**)

$$P(\text{skin}|R) = \frac{P(R|\text{skin})\ P(\text{skin})}{P(R)}$$

what we want
(**posterior**)

normalization term

$$P(R) = P(R|\text{skin})P(\text{skin}) + P(R|\sim\text{skin})P(\sim\text{skin})$$

- What should we use for the prior *P(skin)*?

Bayesian estimation

what we measure (**likelihood**)

domain knowledge (**prior**)

$$P(\text{skin}|R) = \frac{P(R|\text{skin})\ P(\text{skin})}{P(R)}$$

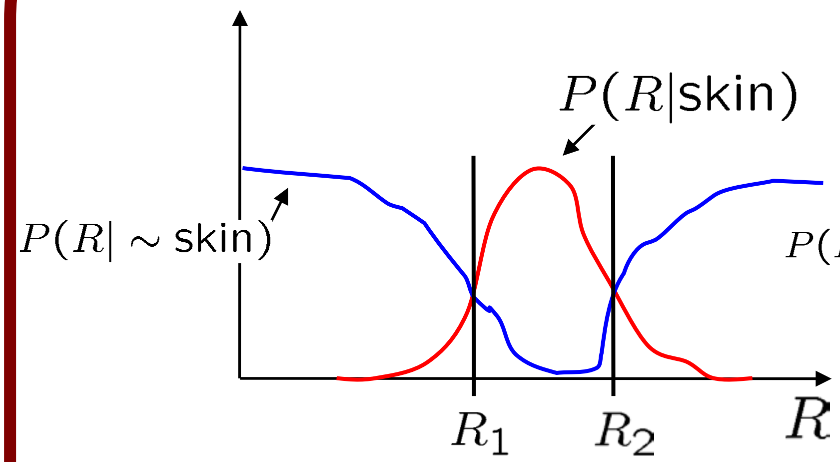
what we want (**posterior**)

normalization term

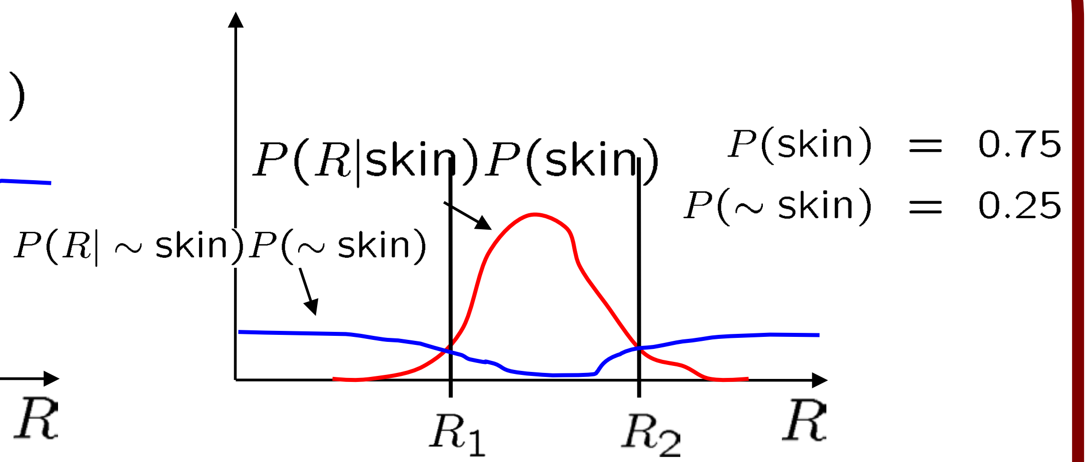
$$P(R) = P(R|\text{skin})P(\text{skin}) + P(R|\sim\text{skin})P(\sim\text{skin})$$

- What could we use for the prior $P(skin)$?
  - Could use domain knowledge
  - $P(skin)$ may be larger if we know the image contains a person
  - for a portrait, $P(skin)$ may be higher for pixels in the center
- Could learn the prior from the training set. How?
  - P(skin) may be proportion of skin pixels in training set

Bayesian estimation



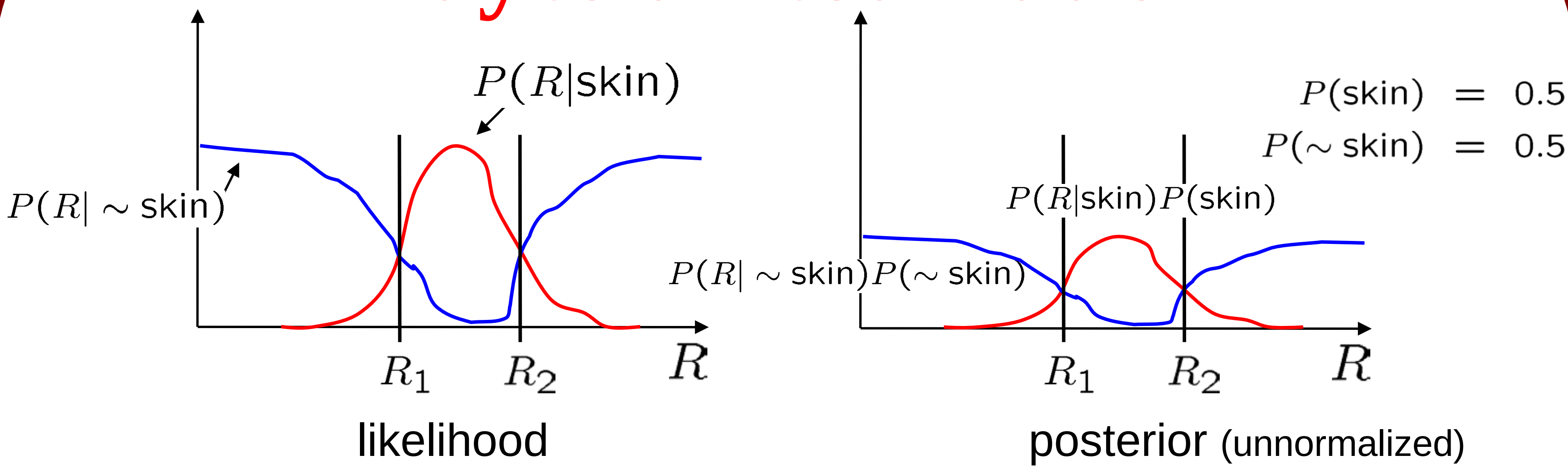
likelihood



posterior (unnormalized)

- Bayesian estimation = minimize probability of misclassification
  - Goal is to choose the label (skin or ~skin) that maximizes the posterior
    - this is called **Maximum A Posteriori (MAP) estimation**

Bayesian estimation



- Suppose the prior is uniform: P(skin) = P(~skin) = 0.5
  - in this case $p(skin|R) = cp(R|skin)$ and $p(\sim skin|R) = cp(R|\sim skin)$
  - maximizing the posterior is equivalent to maximizing the likelihood
  - $p(skin|R) > p(\sim skin|R)$ only if $p(R|skin) > p(R|\sim skin)$
    - Maximum A Posteriori (MAP) estimation equals Maximum Likelihood estimation