Lecture 2

Classification: Bayesian methods

Classification examples

focus for now on *classification* problems

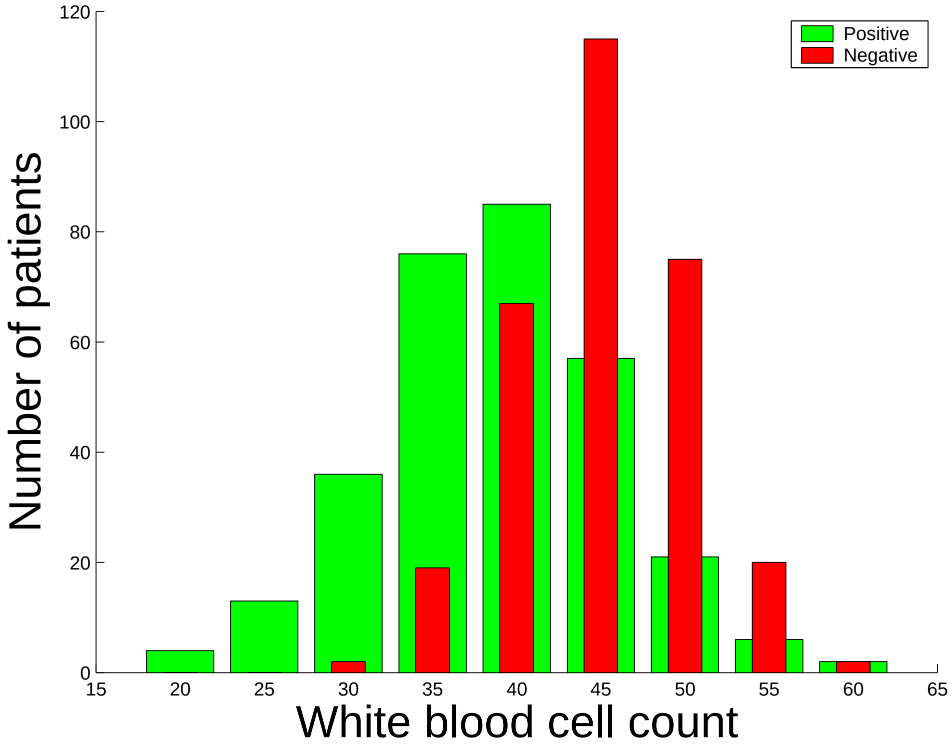
examples:

- Is burglar in house given that alarm just went off?
- Is opponent likely to make this move now?
- Does this patient have cancer?

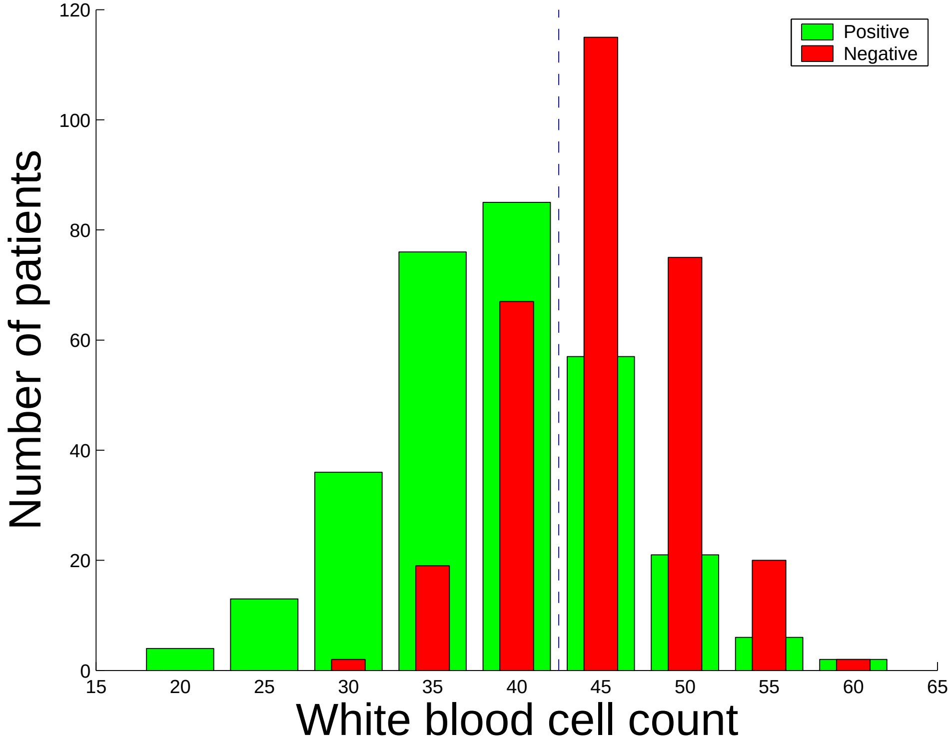
classify on the basis of information extracted from training samples

diagnosis example: doctor deciding if patient has diabetes

Diabetes example

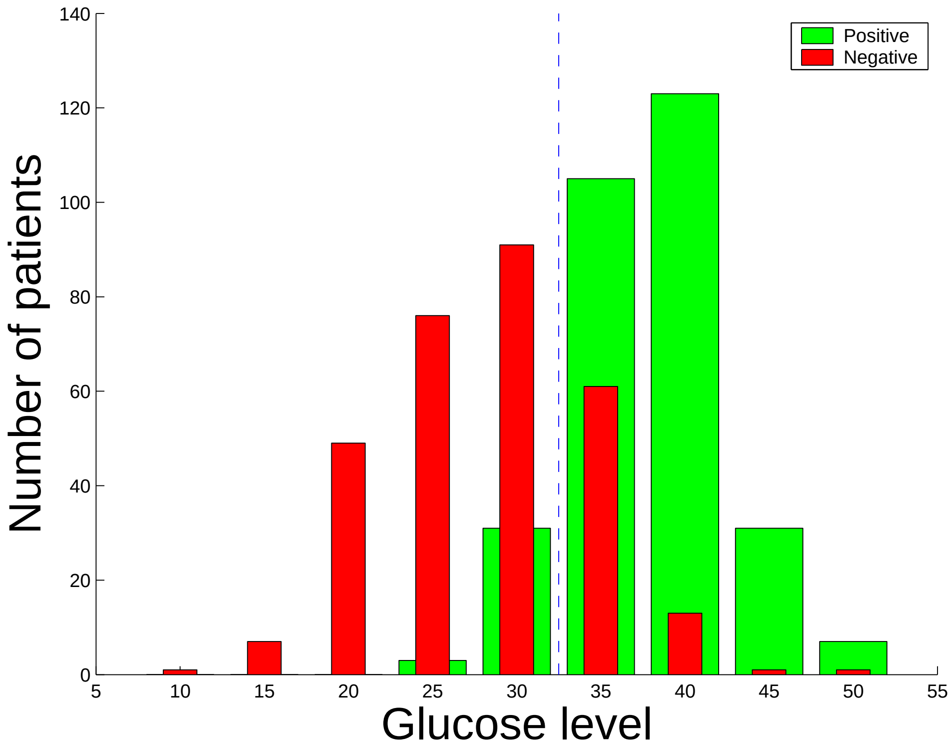


Diabetes example: Decision boundary

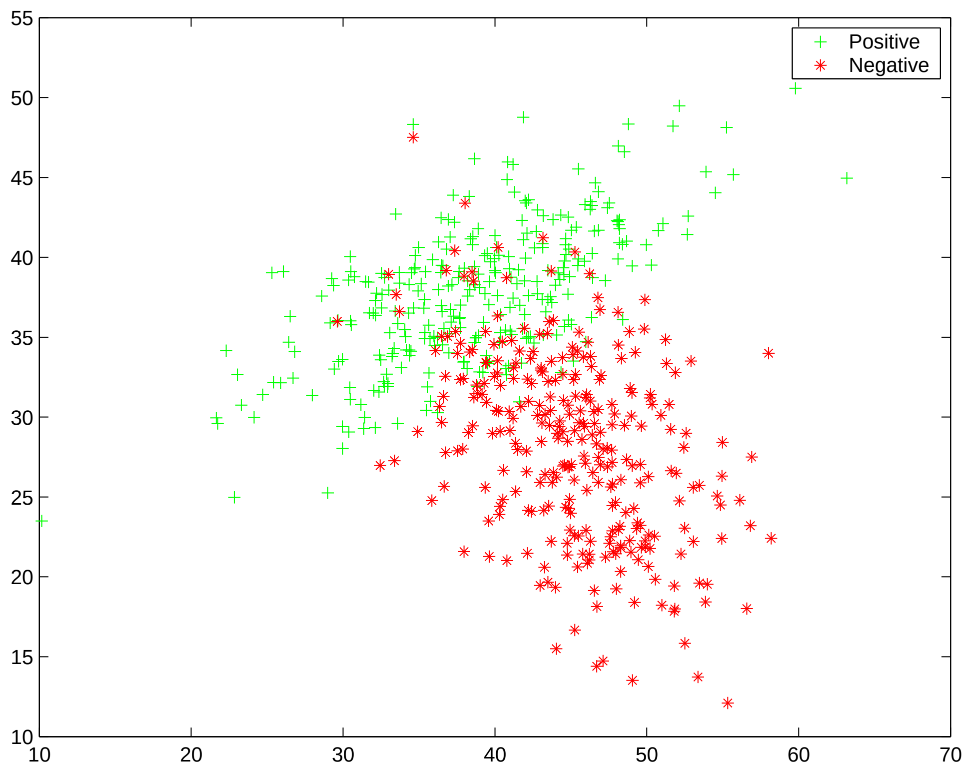


decision boundary used to classify unseen test examples (new patients)

Diabetes example: Better feature?



Diabetes example: Two features



Decision theory

Can classify by counting in bins, or by forming decision boundary

How complicated should the classifer be?

Aim to maximize *generalization*: correct classification of unseen test examples

Must make assumptions about domain, decision rule

Possibly other factors enter into decision: may be more costly to decide that someone does not have diabetes when they in fact do

Decision theory: make a decision to minimize cost

Probabilistic approach, notation

Given that knowledge about domain is incomplete, need to formulate degree of belief; apply probability methods

Prior (unconditional) probability: $P(Roll = 3) = 1/6$

Roll is *random variable* – types include

- boolean (1/0; T/F): $P(\text{Heads} = \text{True}) = .5 = 1 - P(\neg Heads)$
- multinomial (discrete values): $P(Roll = 6)$
- continuous: $p(TempTomorrow = 34\,\text{deg})$

Probability distribution: probabilities associated with all possible values of random variable:

$$\mathbf{P}(R) = < .1, .1, .1, .1, .1, .5 >$$

Joint probability: probability of combination of values of random variables:

$$P(R_1 = 6, R_2 = 6) = P(R_1 = 6 \wedge R_2 = 6) = 1/36$$

Conditional Probability

conditional probability $P(A|B)$: basic expressions in Bayesian formalism for probabilities

"probability of A given that all we know is B"

posterior – conditioned on evidence - given that B is known with certainty

Bayes Rule:

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

follows directly from *product rule*: $P(A|B)P(B) = P(A,B) = P(B|A)P(A)$

independent variables: $P(A|B) = P(A)$;
e.g., $P(Rain|R=6) = P(Rain)$

Bayesian classification

Apply Bayes Rule: c is the class, $\{v\}$ observed attribute values:

$$P(c|\{v\}) = \frac{P(\{v\}|c)P(c)}{P(\{v\})}$$

If we assume K possible disjoint diagnoses, $c_1, .., c_K$

$$P(c_k|\{v\}) = \frac{P(c_k)P(\{v\}|c_k)}{P(\{v\})}$$

$P(\{v\})$ may not be known, but total probability of diagnoses is 1

$P(\{v\})$ (the *evidence*): $\sum_k \frac{P(c_k)P(\{v\}|c_k)}{P(\{v\})} = 1$

$$\Rightarrow P(\{v\}) = \sum_k P(c_k)P(\{v\}|c_k)$$

Need to know $P(c_k), P(\{v\}|c_k)$ for all k

Bayes Rule: $posterior = \frac{likelihood*prior}{evidence}$

Bayesian classification: MAP vs. ML

rather than computing full posterior, can simplify computation if interested in classification

1. ML (Maximum Likelihood) Hypothesis

   assume all hypotheses equiprobable a priori – simply maximize *data likelihood*:

   $$c_{ML} = \arg\max_{c \in C} P(\{v\}|c)$$

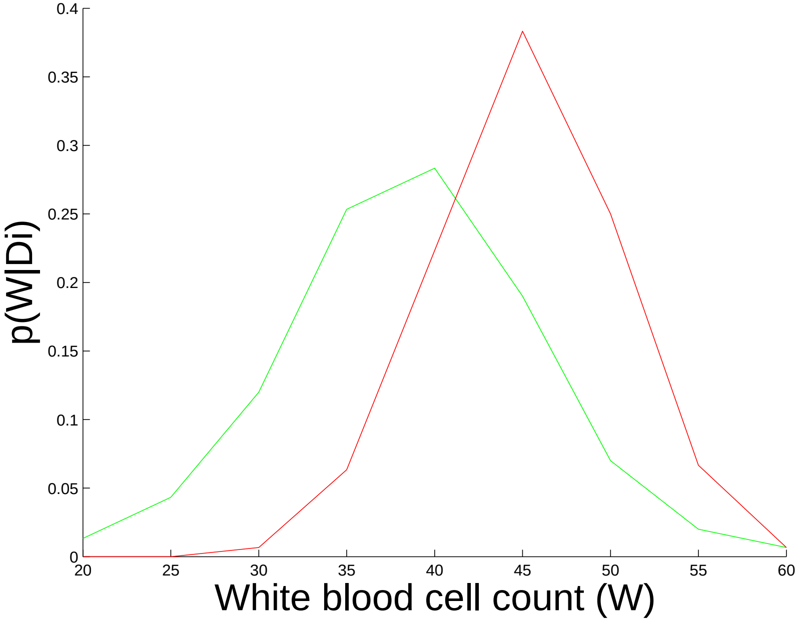
2. MAP (Maximum A Posteriori) Class Hypothesis

   $$\begin{aligned} c_{MAP} &= \arg\max_{c \in C} P(c|\{v\}) \\ &= \arg\max_{c \in C} \frac{P(\{v\}|c)P(c)}{P(\{v\})} \end{aligned}$$

   can ignore denominator because same for all c

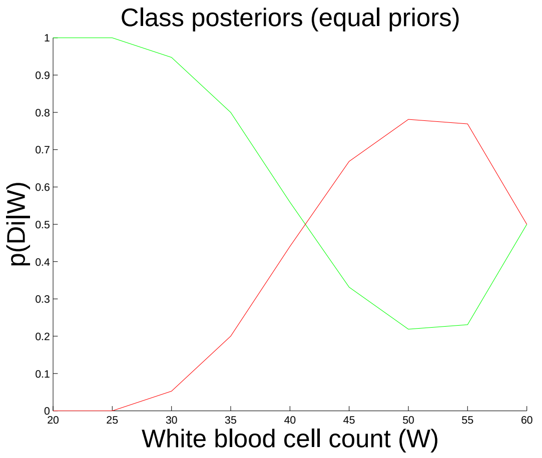
Bayes Theorem: Example

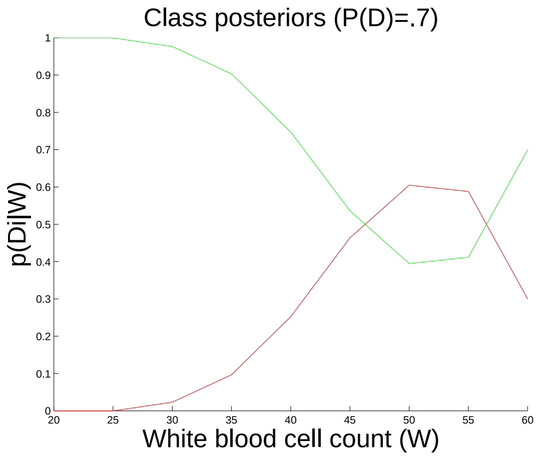
use training examples to estimate class-conditional probability density functions for white-blood cell count (W)

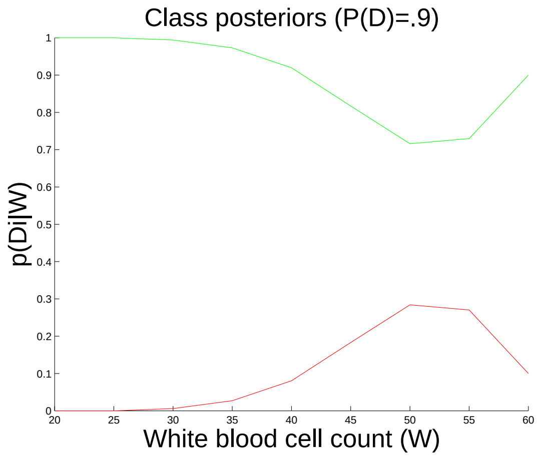


Could use these to select maximum likelihood hypothesis

Suppose most patients in database have diabetes (class priors)







need to form density from samples:

- sort data based on class label c
- estimate $P(c)$ by counting
- estimate $P(\{v\}|c)$ separately within each class

issues: smoothing? assume form of density?