# Object Recognition: Conceptual Issues

Slides adapted from Fei-Fei Li, Rob Fergus, Antonio Torralba, and K. Grauman

# Issues in recognition

The statistical viewpoint

Generative vs. discriminative methods

Model representation

Generalization, bias vs. variance

Supervision

Datasets

# Discrete Random Variables

*X* denotes a random variable.

*X* can take on a countable number of values in $\{x_1, x_2, \ldots, x_n\}$.

$P(X=x_i)$, or $P(x_i)$, is the probability that the random variable *X* takes on value $x_i$.

$P(.)$ · is called probability mass function.

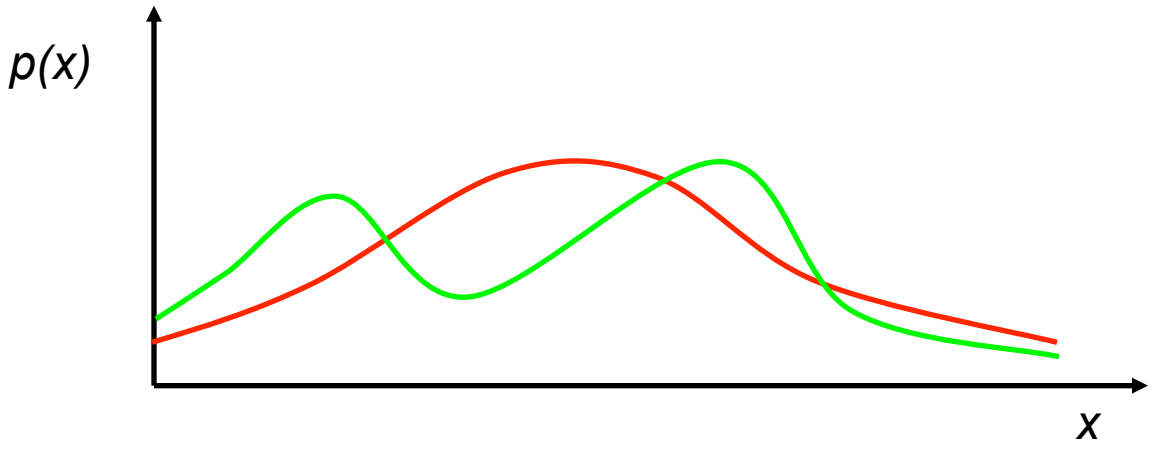
# Continuous Random Variables

*X* takes on values in the continuum.

*p(X=x)*, or *p(x)*, is a probability density function.

$$\Pr(x \in (a,b)) = \int_a^b p(x)dx$$

E.g.



# Joint and Conditional Probability

$P(X=x \text{ and } Y=y) = P(x,y)$

If X and Y are independent then

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

$P(x \mid y)$ is the probability of $x$ given $y$

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

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

If X and Y are independent then

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

# Law of Total Probability, Marginals

**Discrete case**

$$\sum_x P(x) = 1$$

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

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

**Continuous case**

$$\int p(x)\, dx = 1$$

$$p(x) = \int p(x, y)\, dy$$

$$p(x) = \int p(x \mid y)\, p(y)\, dy$$

# Bayes Formula

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

$$\Rightarrow$$

$$P(x \mid y) = \frac{P(y \mid x)\ P(x)}{P(y)} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

# Normalization

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

$$\eta = P(y)^{-1} = \frac{1}{\sum_x P(y \mid x) P(x)}$$

Algorithm:

$$\forall x : \mathrm{aux}_{x|y} = P(y \mid x)\ P(x)$$

$$\eta = \frac{1}{\sum_x \mathrm{aux}_{x|y}}$$

$$\forall x : P(x \mid y) = \eta\ \mathrm{aux}_{x|y}$$

# Bayes' Rule

Bayes Rule for point probabilities

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

or in distribution form

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Useful for assessing diagnostic probability from causal probability:

P(Cause|Effect) = P(Effect|Cause) P(Cause)/P(Effect)
E.g., let *M* be meningitis, *S* be stiff neck:

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.8 \times 0.001}{0.1} = 0.008$$

# Bayes' Rule

Bayes Rule for point probabilities

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.8 \times 0.001}{0.1} = 0.008$$

$$P(s) = P(s \mid m)P(m) + P(s \mid \neg m)P(\neg m)$$

Here we assumed that prior probability is known quantity – exactly 0.001 (point estimate)

Bayesian approach capture the uncertainty about prior as a distribution

Then the posterior will also be distribution

# Simple Example of Recognition

Suppose a we obtain measurement *z*

What is *P(zebra|z)?*

# Causal vs. Diagnostic Reasoning

*P(zebra|z)* is diagnostic.

*P(z|zebra)* is causal.

Often causal knowledge is easier to obtain.

Bayes rule allows us to use causal knowledge:

**count frequencies!**

$$P(zebra \mid z) = \frac{P(z \mid zebra)P(zebra)}{P(z)}$$

# Combining Evidence

Suppose we obtain another observation $z_2$.

How can we integrate this new information?

More generally, how can we estimate $P(x| z_1...z_n)$?

# Example: Second Measurement

$P(z_2|zebra) = 0.5$ $\quad\quad$ $P(z_2|\neg zebra) = 0.6$

$P(zebra|z_1)=2/3$

$$P(zebra \mid z_2, z_1) = \frac{P(z_2 \mid zebra)\, P(zebra \mid z_1)}{P(z_2 \mid zebra)\, P(zebra \mid z_1) + P(z_2 \mid \neg zebra)\, P(\neg zebra \mid z_1)}$$

$$= \frac{\frac{1}{2} \cdot \frac{2}{3}}{\frac{1}{2} \cdot \frac{2}{3} + \frac{3}{5} \cdot \frac{1}{3}} = \frac{5}{8} = 0.625$$

- $z_2$ lowers the probability that the picture is zebra.

# Object categorization: the statistical viewpoint

- MAP decision: $p(zebra \mid image)$

vs.

$p(no\ zebra \mid image)$

$$P(x \mid y) = \frac{P(y \mid x)\ P(x)}{P(y)} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

# Object categorization: the statistical viewpoint

- MAP decision: $p(zebra \mid image)$

vs.

$p(no\ zebra \mid image)$

- Bayes rule:

$$\underbrace{p(zebra \mid image)}_{\text{posterior}} \propto \underbrace{p(image \mid zebra)}_{\text{likelihood}} \underbrace{p(zebra)}_{\text{prior}}$$

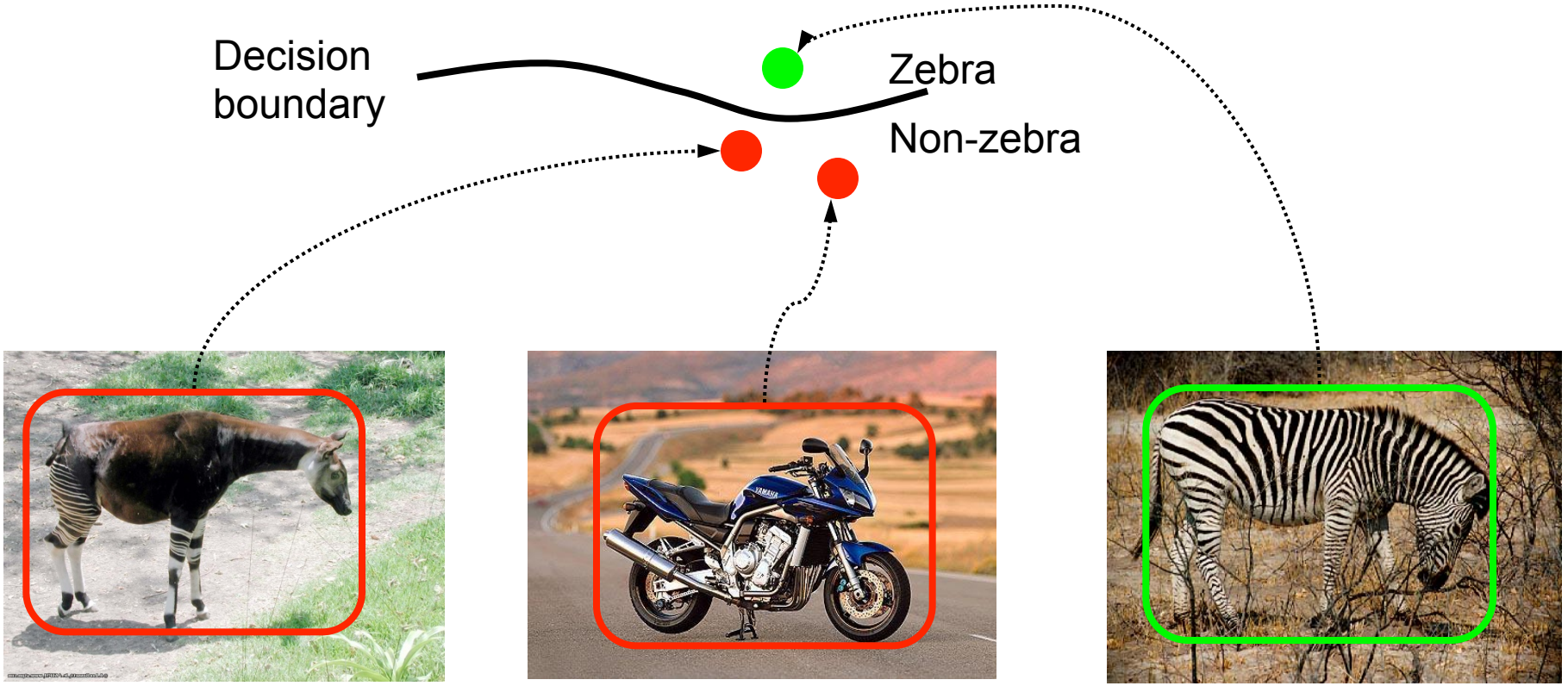
# Object categorization: the statistical viewpoint

$$\underbrace{p(zebra \mid image)}_{\text{posterior}} \propto \underbrace{p(image \mid zebra)}_{\text{likelihood}} \underbrace{p(zebra)}_{\text{prior}}$$

- **Discriminative methods:** **model posterior**
- **Generative methods:** **model likelihood and prior**

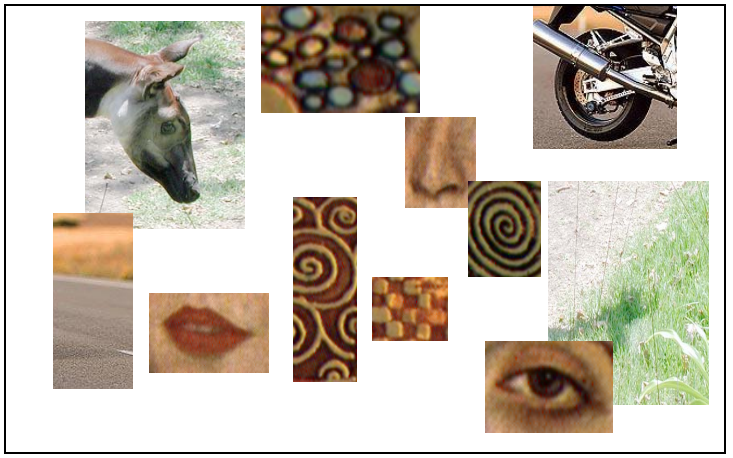
# Discriminative methods

- Direct modeling of $p(zebra \mid image)$



# Generative methods

- Model $p(image \mid zebra)$ and $p(image \mid no\ zebra)$



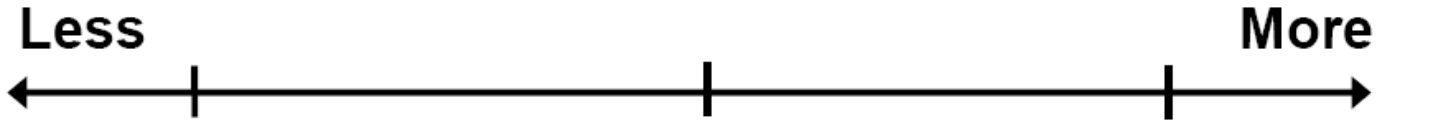
| $p(image \mid zebra)$ | $p(image \mid no\ zebra)$ |
|---|---|
| Low | Middle |
| High | Middle→Low |

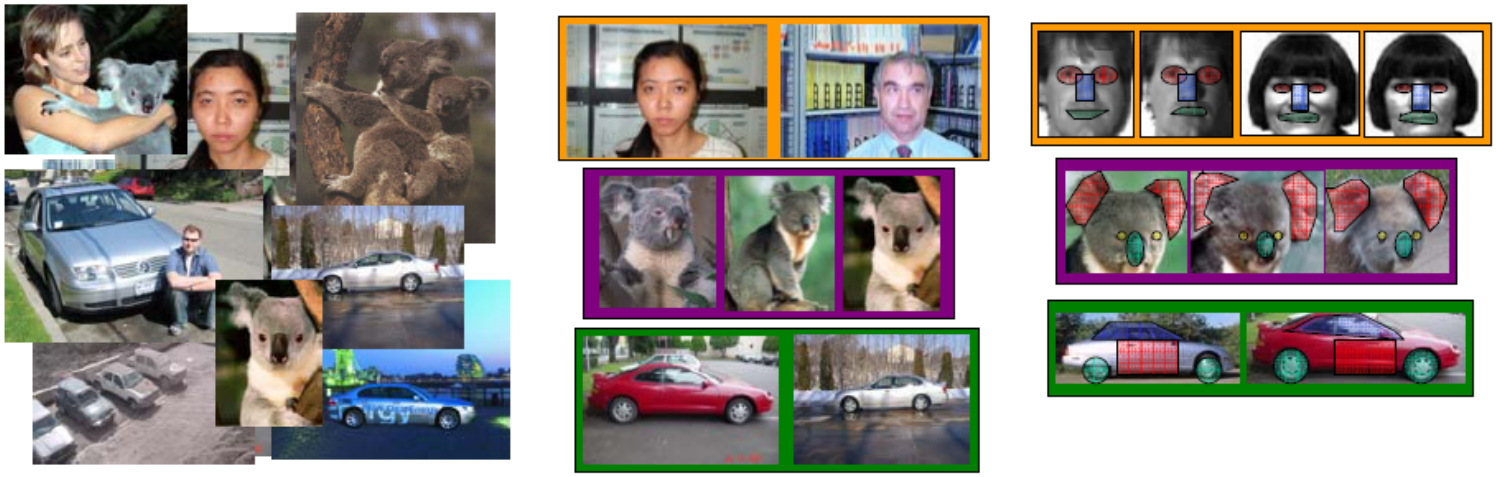
# Generative vs. discriminative methods

- Generative methods
  - + Interpretable
  - + Can be learned using images from just a single category
  - – Sometimes we don't need to model the likelihood when all we want is to make a decision

- Discriminative methods
  - + Efficient
  - + Often produce better classification rates
  - – Can be hard to interpret
  - – Require positive and negative training data

# Spectrum of supervision



Learning approaches proceed in supervised way: need some labeled data

Unsupervised

“Weakly” supervised

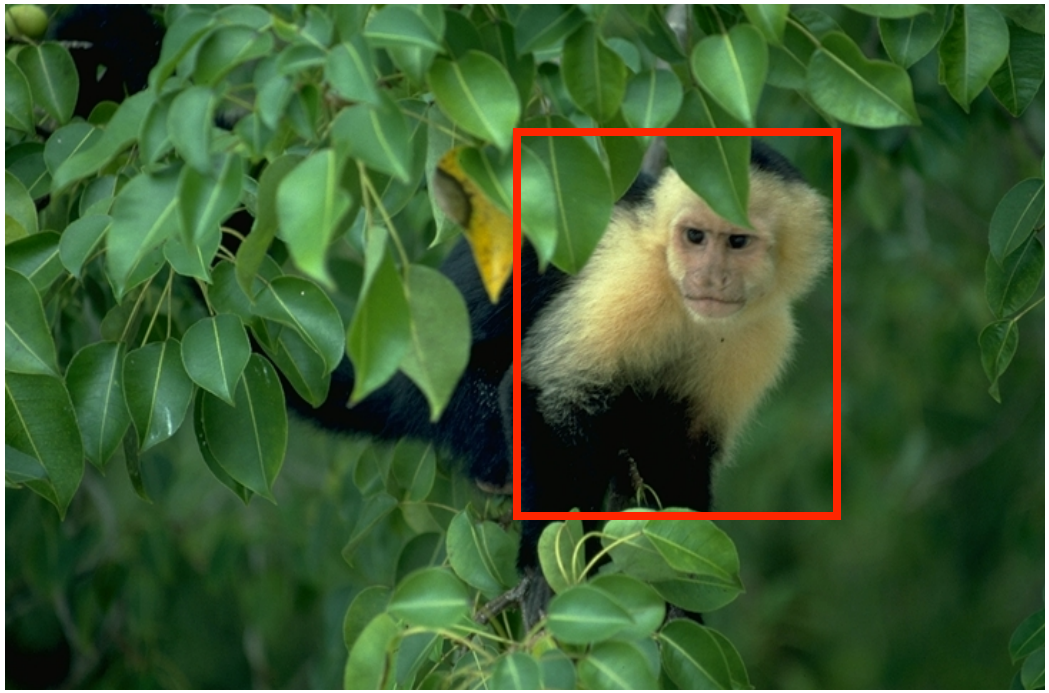
Supervised

Definition depends on task

# What task?

- Classification
  - Object present/absent in image
  - Background may be correlated with object

- Localization / Detection
  - Localize object within the frame
  - Bounding box or pixel-level segmentation



# Caltech 101 & 256

http://www.vision.caltech.edu/Image_Datasets/Caltech101/
http://www.vision.caltech.edu/Image_Datasets/Caltech256/

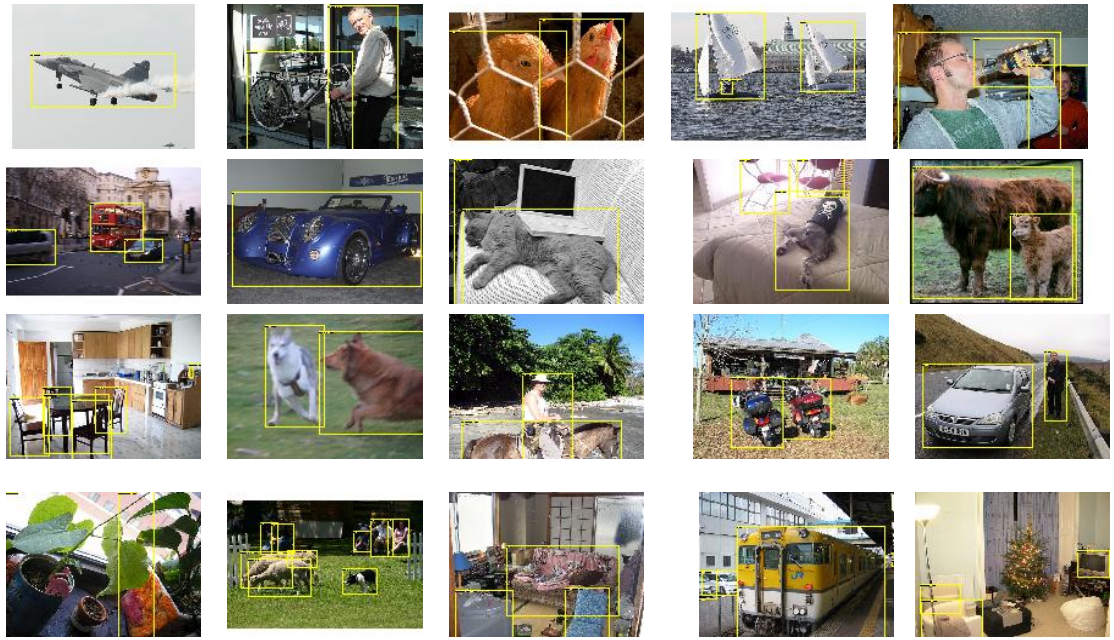
Fei-Fei, Fergus, Perona, 2004

Griffin, Holub, Perona, 2007

# The PASCAL Visual Object Classes Challenge (2005-2009)

http://pascallin.ecs.soton.ac.uk/challenges/VOC/

**2008 Challenge classes:**
*Person:* person
*Animal:* bird, cat, cow, dog, horse, sheep
*Vehicle:* aeroplane, bicycle, boat, bus, car, motorbike, train
*Indoor:* bottle, chair, dining table, potted plant, sofa, tv/monitor



# The PASCAL Visual Object Classes Challenge (2005-2009)

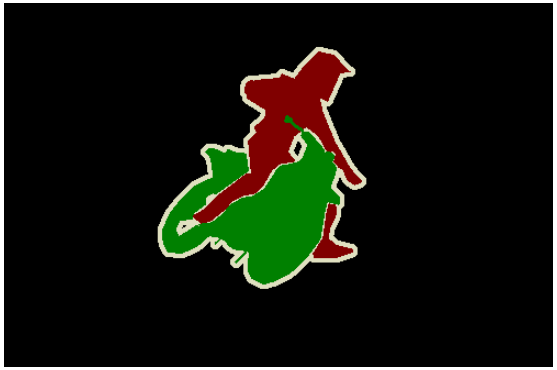
http://pascallin.ecs.soton.ac.uk/challenges/VOC/

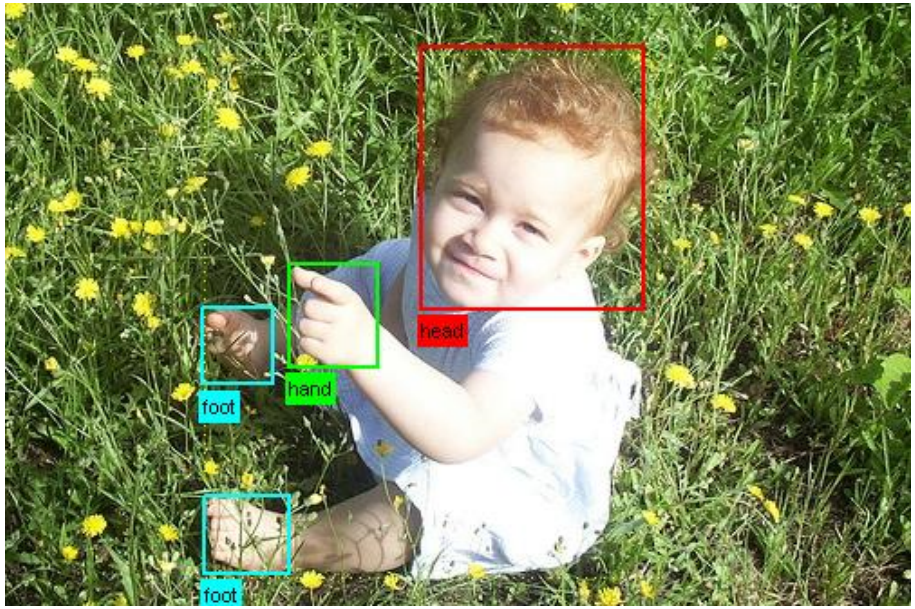
- Main competitions
  - **Classification:** For each of the twenty classes, predicting presence/absence of an example of that class in the test image
  - **Detection:** Predicting the bounding box and label of each object from the twenty target classes in the test image

# The PASCAL Visual Object Classes Challenge (2005-2009)

http://pascallin.ecs.soton.ac.uk/challenges/VOC/

- "Taster" challenges
  - **Segmentation:** Generating pixel-wise segmentations giving the class of the object visible at each pixel, or "background" otherwise
  - **Person layout:** Predicting the bounding box and label of each part of a person (head, hands, feet)





# Labeling with games

http://www.gwap.com/gwap/

*Figure 1. Partners agreeing on an image in the* ESP Game. *Neither player can see the other's guesses.*

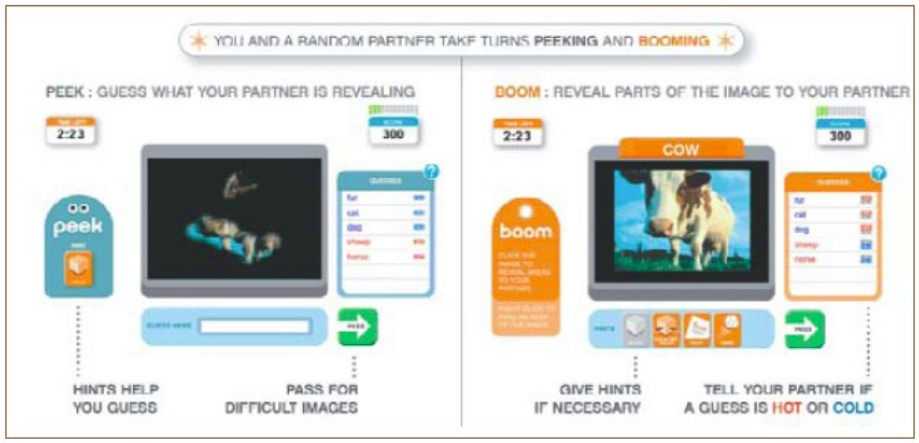



*Figure 2.* Peekaboom. *"Peek" tries to guess the word associated with an image slowly revealed by "Boom."*

L. von Ahn, L. Dabbish, 2004; L. von Ahn, R. Liu and M. Blum, 2006

# LabelMe

http://labelme.csail.mit.edu/

Russell, Torralba, Murphy, Freeman, 2008

# Summary

- Recognition is the “grand challenge” of computer vision
- History
  - Geometric methods
  - Appearance-based methods
  - Sliding window approaches
  - Local features
  - Parts-and-shape approaches
  - Bag-of-features approaches
- Issues
  - Generative vs. discriminative models
  - Supervised vs. unsupervised methods
  - Tasks, datasets

# Discriminative Methods

- Object/scene category recognition
- Multi-class classification problem
- Supervised setting
- Given examples of images with category labels
- Learn classifier to predict labels of new images

# Discriminative methods

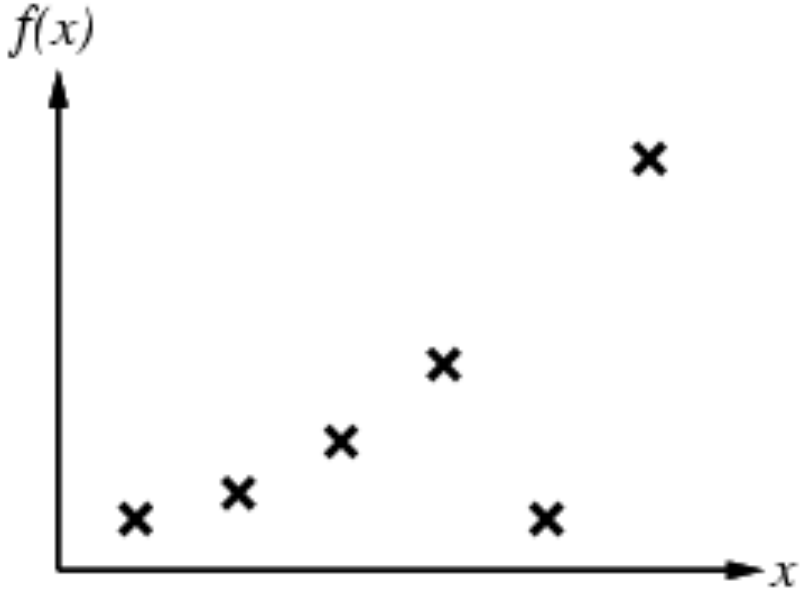
- Brief overview of machine learning
- Linear regression
- Logistic regression
- Decision trees
- Boosting idea
- SVM

*Slides from Russell and Norvig, AI book*

# Inductive learning method

- Construct/adjust *h* to agree with *f* on training set
- (*h* is consistent if it agrees with *f* on all examples)

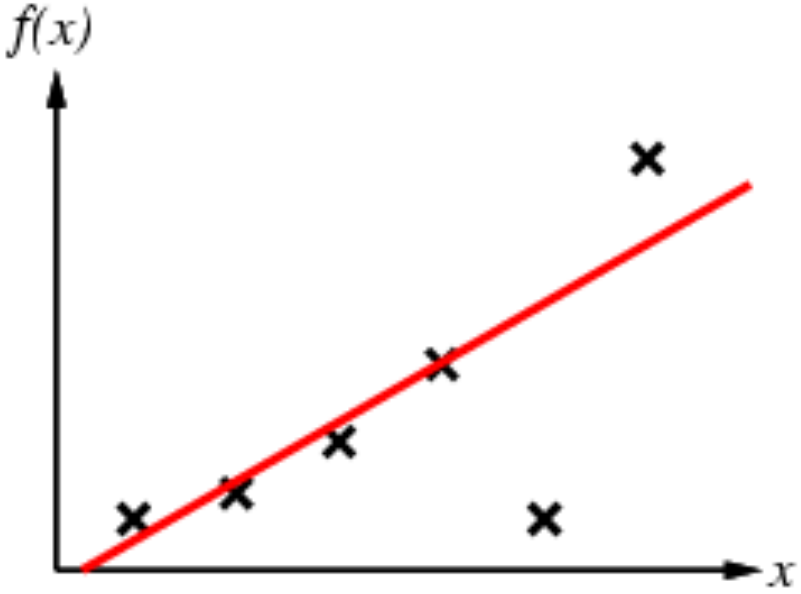
- E.g., curve fitting:



# Inductive learning method

- Construct/adjust *h* to agree with *f* on training set
- (*h* is consistent if it agrees with *f* on all examples)

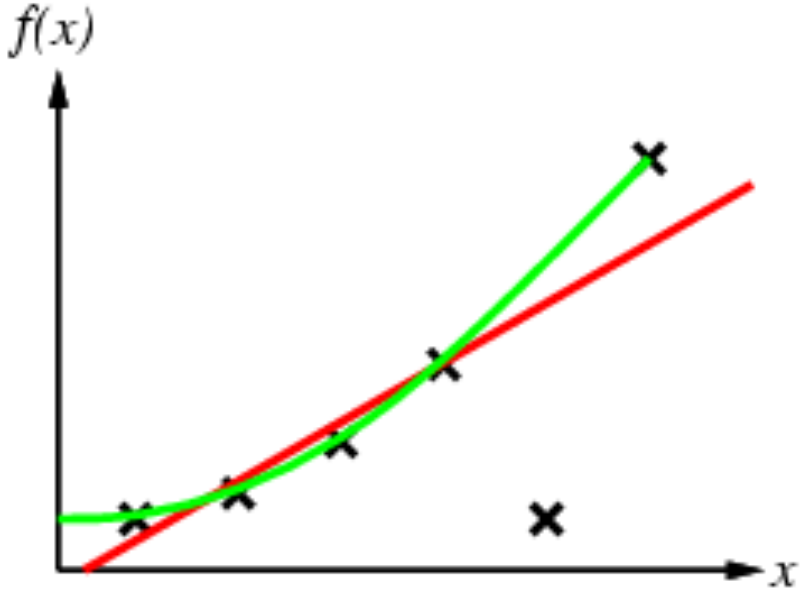
- E.g., curve fitting:



# Inductive learning method

- Construct/adjust *h* to agree with *f* on training set
- (*h* is consistent if it agrees with *f* on all examples)

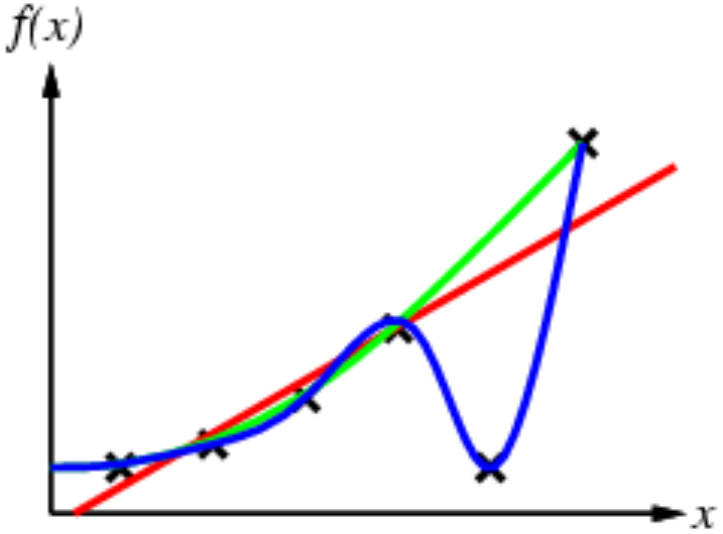
- E.g., curve fitting:



# Inductive learning method

- Construct/adjust *h* to agree with *f* on training set
- (*h* is consistent if it agrees with *f* on all examples)

- E.g., curve fitting:



# Linear regression

- Fit function to the data so you can predict future values
- Given data *(x1,y1), (x2, y2), … (xn, yn)*
- Find such hypothesis f, such that y = f(x) has small error of future data
- Choose the form of function and estimate the data using linear least squares techniques (in closed form, or gradient descent)

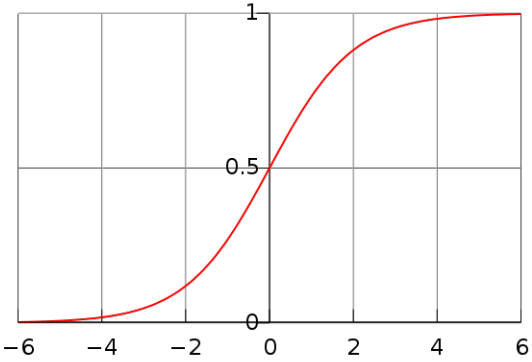
# Logistic Regression

- Similar as linear regression, but now *y's* are only +/- or 0/1
- denoting positive or negative examples of a class
- We know that our function should have values 0 or 1
- Transform to continuous setting – construct such as *h* where function would have values only between 0-1

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

- g(.) is a sigmoid, logistic function

$$g(z) = \frac{1}{1 + e^{-z}}$$




- Property of the derivative

$$g'(z) = g(z)(1 - g(z))$$

# Logistic Regression

- How do we find parameter $\theta$
- Find such parameters so as to maximize likelihood of the data assume that

$$\begin{aligned} P(y=1|x;\theta) &= h_\theta(x) \\ P(y=0|x;\theta) &= 1-h_\theta(x) \end{aligned}$$

- More compactly

$$p(y|x;\theta) = (h_\theta)^y(1-h_\theta)^{(1-y)}$$

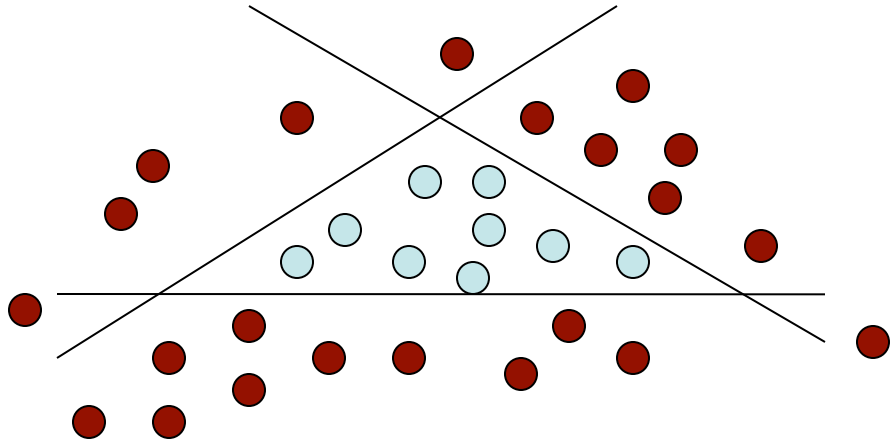
- Resulting gradient ascent rule

$$\theta_j = \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x^{(i)}$$

- Closely related to perceptron learning algorithm where values are forced to be 0-1 - no clear probabilistic interpretation

# Ensemble learning

- Idea: instead of keeping simple hypothesis – keep multiple of them – ensemble of hypotheses
- Increased expressive power
- Enables us to combine several simpler hypotheses to get better prediction

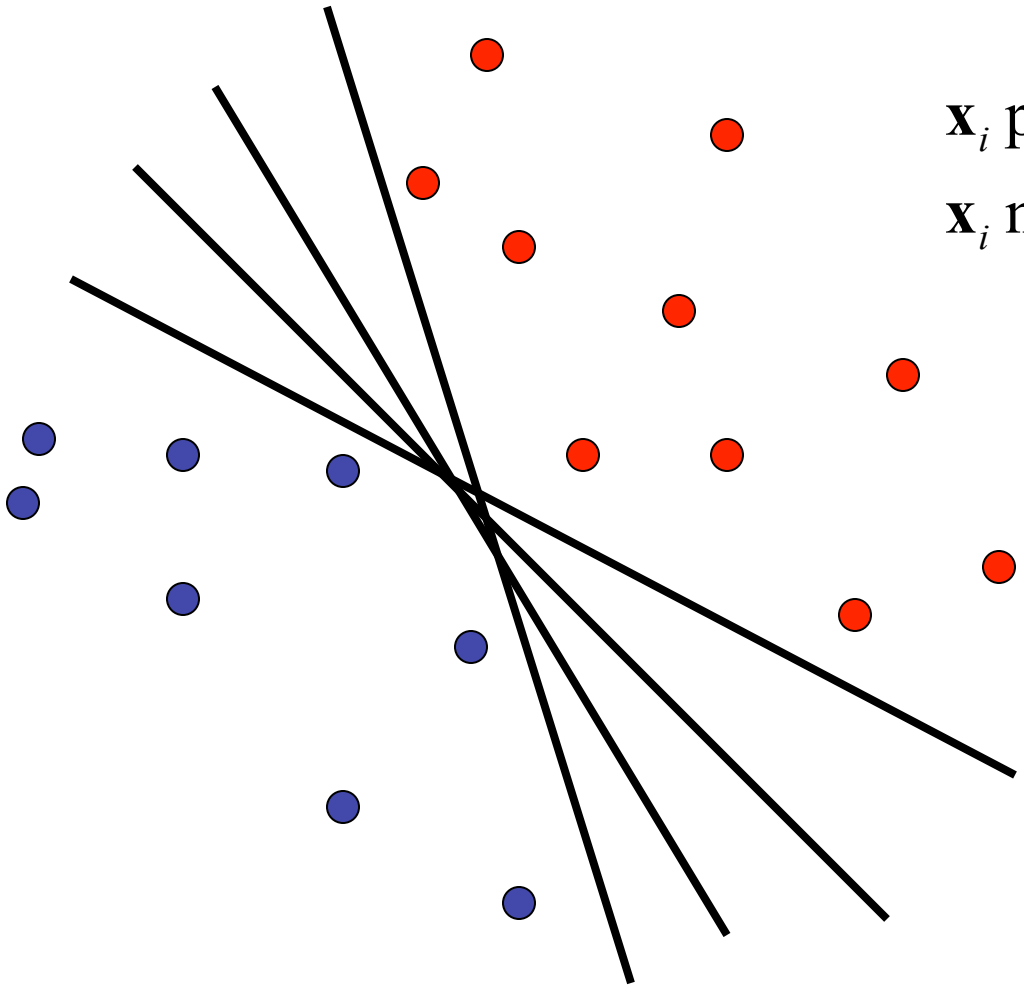


# Ensemble learning

- **Boosting**
- Idea: weighted training set – each training example has an associated weight *w >=0*
- The higher the weight, higher importance of the example in the learning stage
- Start with all examples with weight 1.
- Generate hypothesis *h1*
- *h1* classifies some examples correctly some incorrectly
- Next hypothesis should do better on the missclassified examples – i.e. increase their weights
- Given new weighted training set generate *h2* etc
- We will see examples of these later

# Linear classifiers

- Find linear function (*hyperplane*) to separate positive and negative examples – many such hyperplanes



$$\mathbf{x}_i \text{ positive}: \quad \mathbf{x}_i \cdot \mathbf{w} + b \geq 0$$

$$\mathbf{x}_i \text{ negative}: \quad \mathbf{x}_i \cdot \mathbf{w} + b < 0$$

Which hyperplane is best?

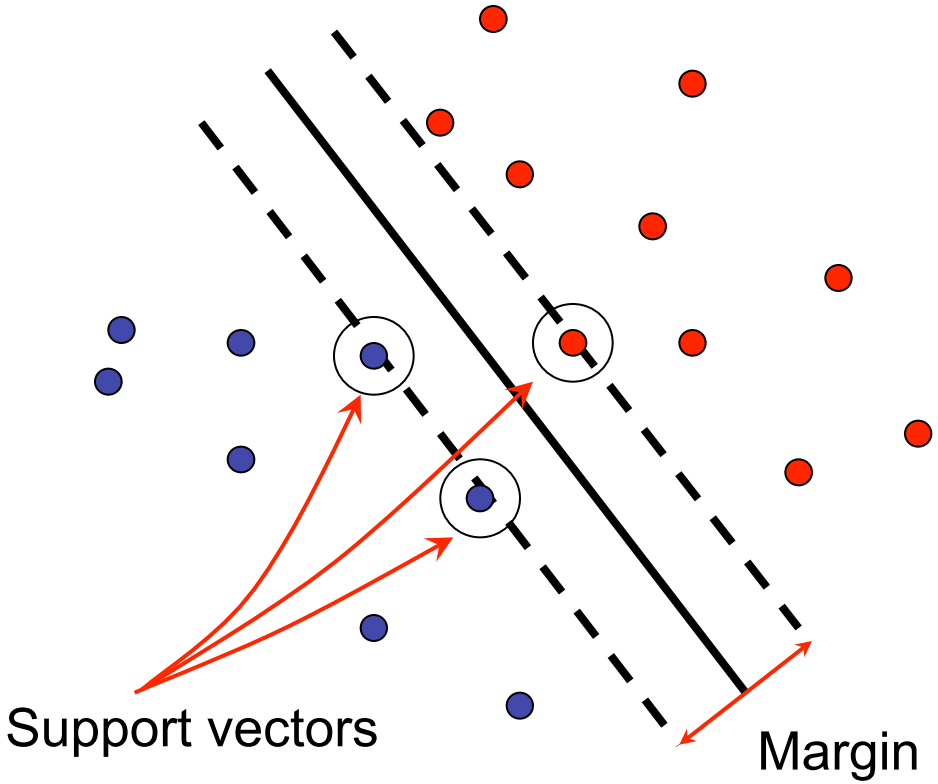
# Support vector machines

- Find hyperplane that maximizes the *margin* between the positive and negative examples

C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 1998

# Support vector machines

- Find hyperplane that maximizes the *margin* between the positive and negative examples




$\mathbf{x}_i$ positive $(y_i = 1)$: $\mathbf{x}_i \cdot \mathbf{w} + b \geq 1$

$\mathbf{x}_i$ negative $(y_i = -1)$: $\mathbf{x}_i \cdot \mathbf{w} + b \leq -1$

For support, vectors, $\mathbf{x}_i \cdot \mathbf{w} + b = \pm 1$

Distance between point and hyperplane: $\dfrac{|\mathbf{x}_i \cdot \mathbf{w} + b|}{\|\mathbf{w}\|}$

Therefore, the margin is $2 / \|\mathbf{w}\|$

C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 1998

# Summary: Discriminative methods

- Nearest-neighbor and k-nearest-neighbor classifiers
  - L1 distance, $\chi^2$ distance, quadratic distance, Earth Mover's Distance
- Support vector machines
  - Linear classifiers
  - Margin maximization
  - The kernel trick
  - Kernel functions: histogram intersection, generalized Gaussian, pyramid match
  - Multi-class
- Of course, there are many other classifiers out there
  - Neural networks, boosting, decision trees, …