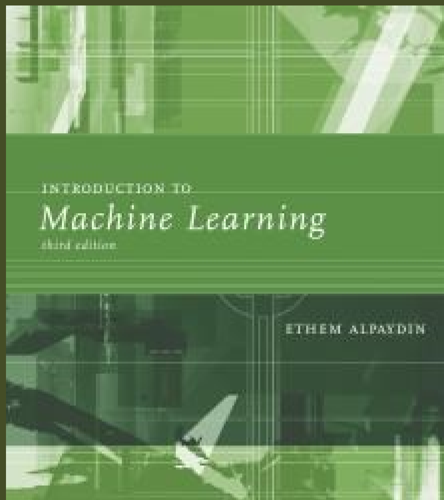


Lecture Slides for

INTRODUCTION TO MACHINE LEARNING

3RD EDİTİON

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml3e

Big Data

- Widespread use of personal computers and wireless communication leads to “big data”
- We are both producers and consumers of data
- Data is not random, it has structure, e.g., customer behavior
- We need “big theory” to extract that structure from data for

  (a) Understanding the process

  (b) Making predictions for the future

Why “Learn” ?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- There is no need to “learn” to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)
  - Solution needs to be adapted to particular cases (user biometrics)

What We Talk About When We Talk About "Learning"

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behavior:

  People who bought "Blink" also bought "Outliers" (www.amazon.com)
- Build a model that is *a good and useful approximation* to the data.

Data Mining

- Retail: Market basket analysis, Customer relationship management (CRM)
- Finance: Credit scoring, fraud detection
- Manufacturing: Control, robotics, troubleshooting
- Medicine: Medical diagnosis
- Telecommunications: Spam filters, intrusion detection
- Bioinformatics: Motifs, alignment
- Web mining: Search engines
- …

What is Machine Learning?

- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
  - Solve the optimization problem
  - Representing and evaluating the model for inference

Applications

- Association
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning (Clustering, Dimensionality Reduction)
- Reinforcement Learning

Learning Associations

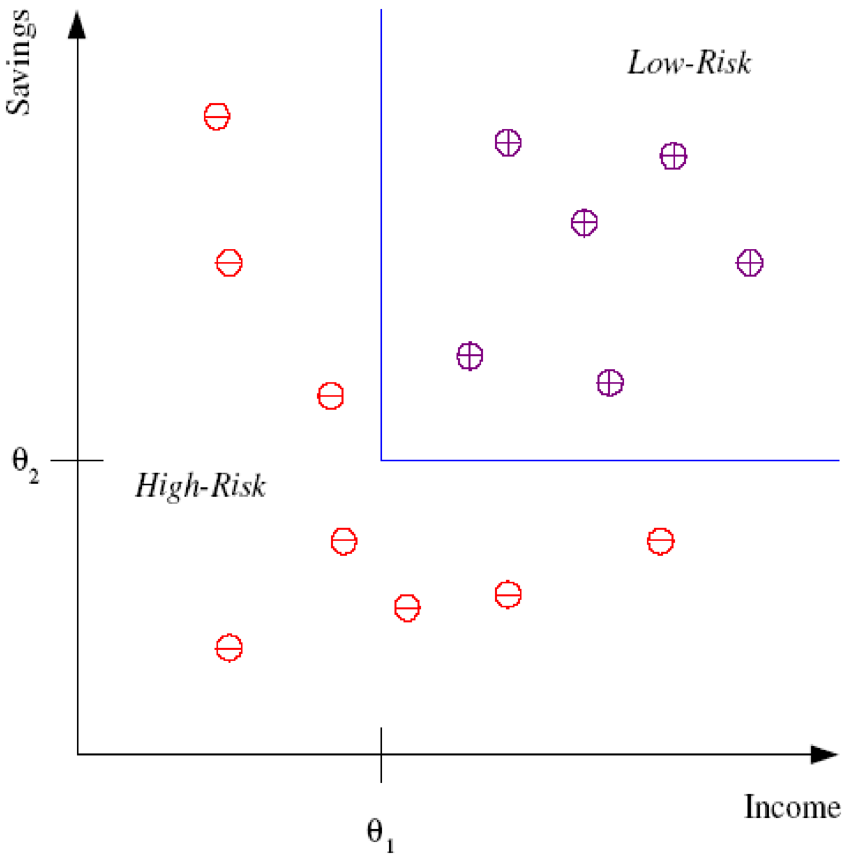
- Basket analysis:

  $P(Y \mid X)$ probability that somebody who buys X also buys Y where X and Y are products/services.

  Example: $P(\text{chips} \mid \text{beer}) = 0.7$

Classification

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*



Discriminant: IF *income* > θ_1 AND *savings* > θ_2 THEN low-risk ELSE high-risk

Classification: Applications

- Aka Pattern recognition
- Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- Character recognition: Different handwriting styles.
- Speech recognition: Temporal dependency.
- Medical diagnosis: From symptoms to illnesses
- Biometrics: Recognition/authentication using physical and/or behavioral characteristics: Face, iris, signature, etc
- Outlier/novelty detection:

Face Recognition

Training examples of a person

Test images

ORL dataset,
AT&T Laboratories, Cambridge UK

Regression

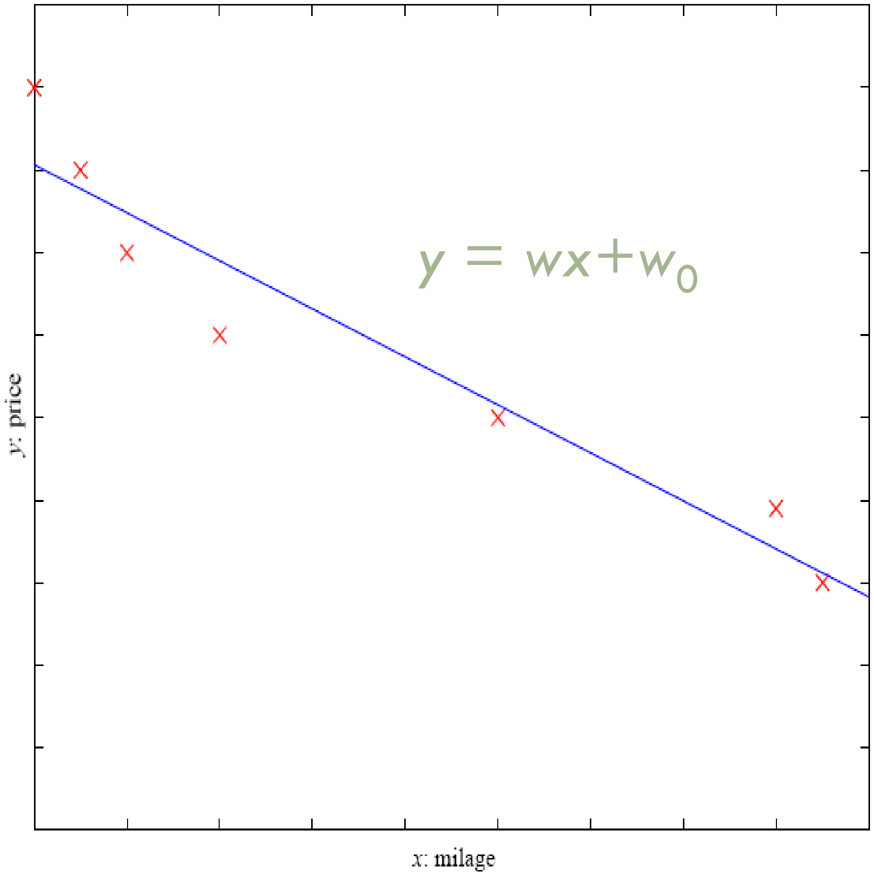
- Example: Price of a used car
- x : car attributes

  y : price

  $$y = g(x \mid \theta)$$

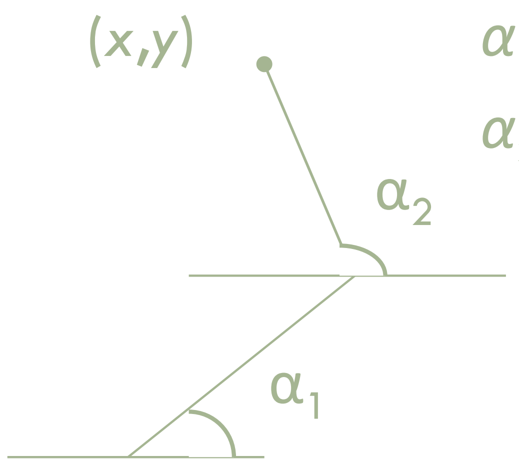
  $g(\,)$ model,

  θ parameters



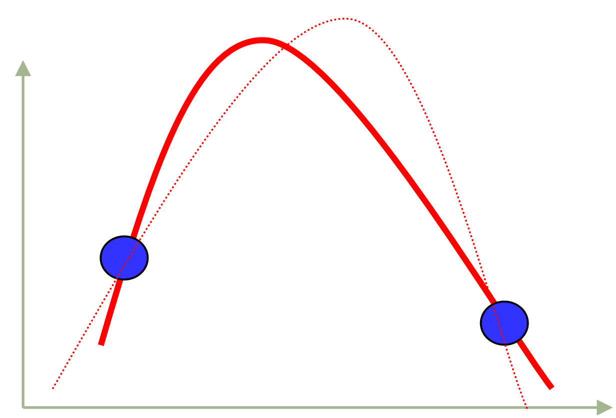
Regression Applications

- Navigating a car: Angle of the steering
- Kinematics of a robot arm



$\alpha_1 = g_1(x,y)$

$\alpha_2 = g_2(x,y)$



- Response surface design

Supervised Learning: Uses

- Prediction of future cases: Use the rule to predict the output for future inputs
- Knowledge extraction: The rule is easy to understand
- Compression: The rule is simpler than the data it explains
- Outlier detection: Exceptions that are not covered by the rule, e.g., fraud

Unsupervised Learning

- Learning "what normally happens"
- No output
- Clustering: Grouping similar instances
- Example applications
  - Customer segmentation in CRM
  - Image compression: Color quantization
  - Bioinformatics: Learning motifs

Reinforcement Learning

- Learning a policy: A sequence of outputs
- No supervised output but delayed reward
- Credit assignment problem
- Game playing
- Robot in a maze
- Multiple agents, partial observability, ...

Resources: Datasets

- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- Statlib: http://lib.stat.cmu.edu/

Resources: Journals

- Journal of Machine Learning Research www.jmlr.org
- Machine Learning
- Neural Computation
- Neural Networks
- IEEE Trans on Neural Networks and Learning Systems
- IEEE Trans on Pattern Analysis and Machine Intelligence
- Journals on Statistics/Data Mining/Signal Processing/Natural Language Processing/Bioinformatics/...

Resources: Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Conference on Artificial Neural Networks (ICANN)
- International Conference on AI & Statistics (AISTATS)
- International Conference on Pattern Recognition (ICPR)
- …

CHAPTER 2:

SUPERVİSED LEARNİNG

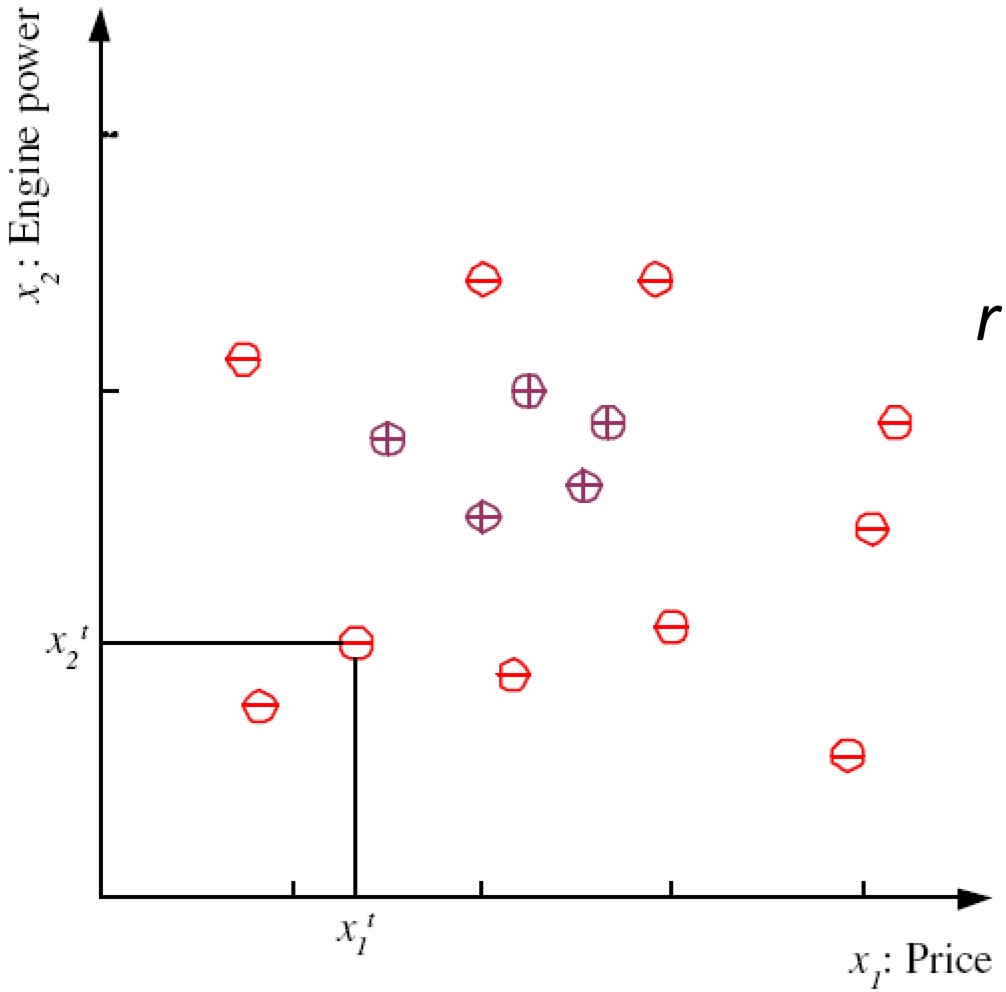
Learning a Class from Examples

- Class C of a "family car"
  - Prediction: Is car x a family car?
  - Knowledge extraction: What do people expect from a family car?
- Output:

  Positive (+) and negative (–) examples
- Input representation:

  x_1: price, x_2 : engine power

Training set $\mathcal{X}$



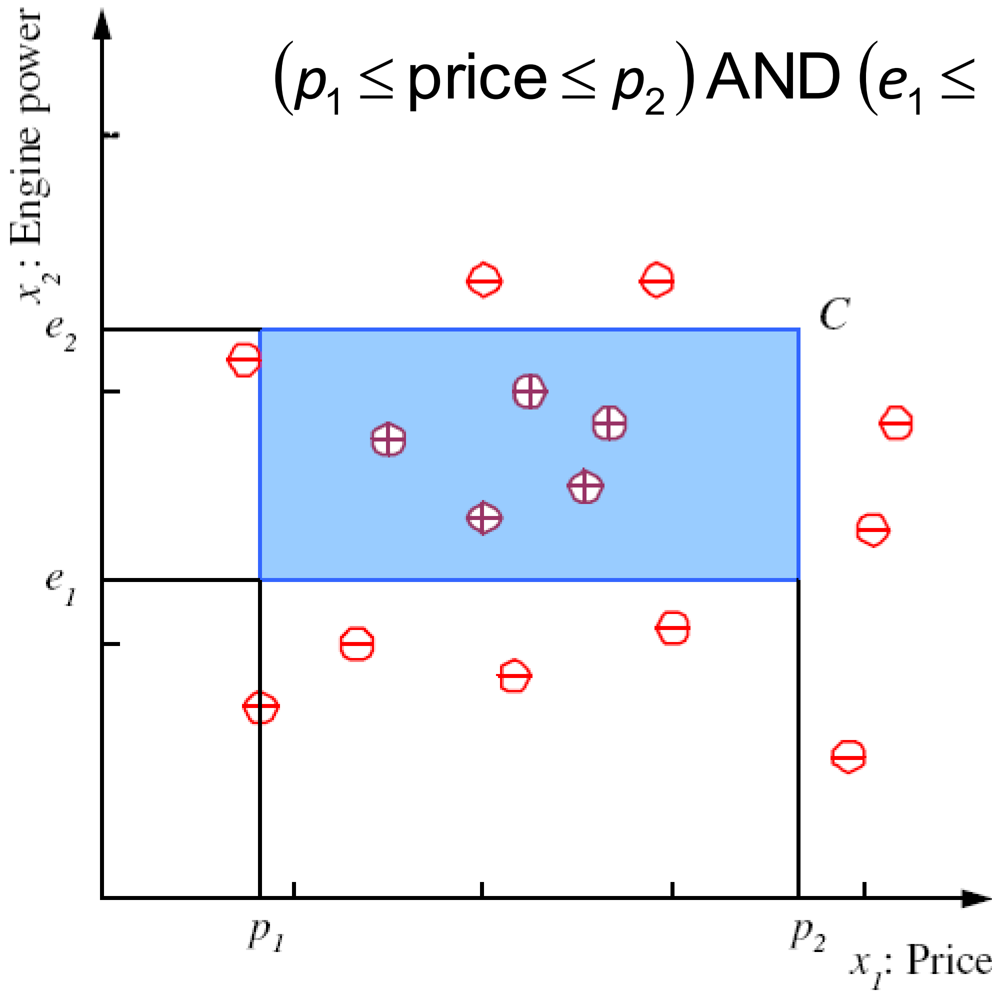
$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^{N}$$

$$r = \begin{cases} 1 \text{ if } \mathbf{x} \text{ is positive} \\ 0 \text{ if } \mathbf{x} \text{ is negative} \end{cases}$$

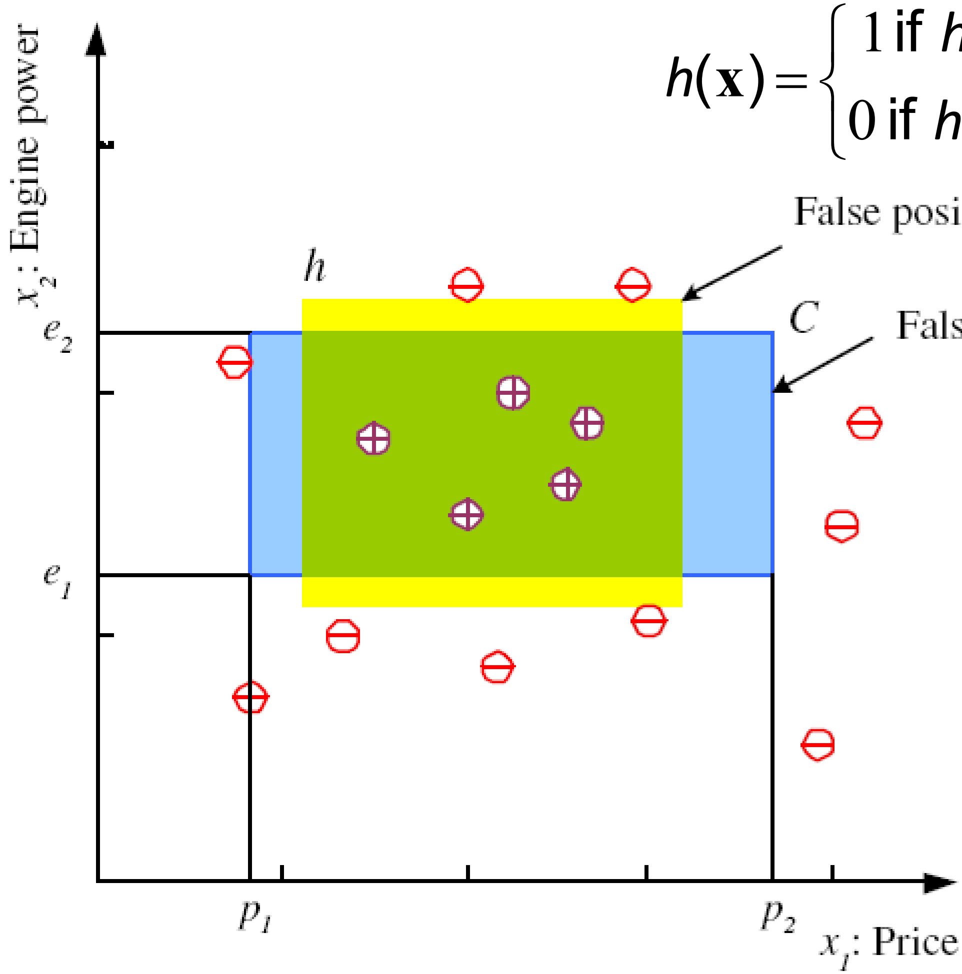
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Class C

$$(p_1 \le \text{price} \le p_2)\ \text{AND}\ (e_1 \le \text{engine power} \le e_2)$$



Hypothesis class $\mathcal{H}$

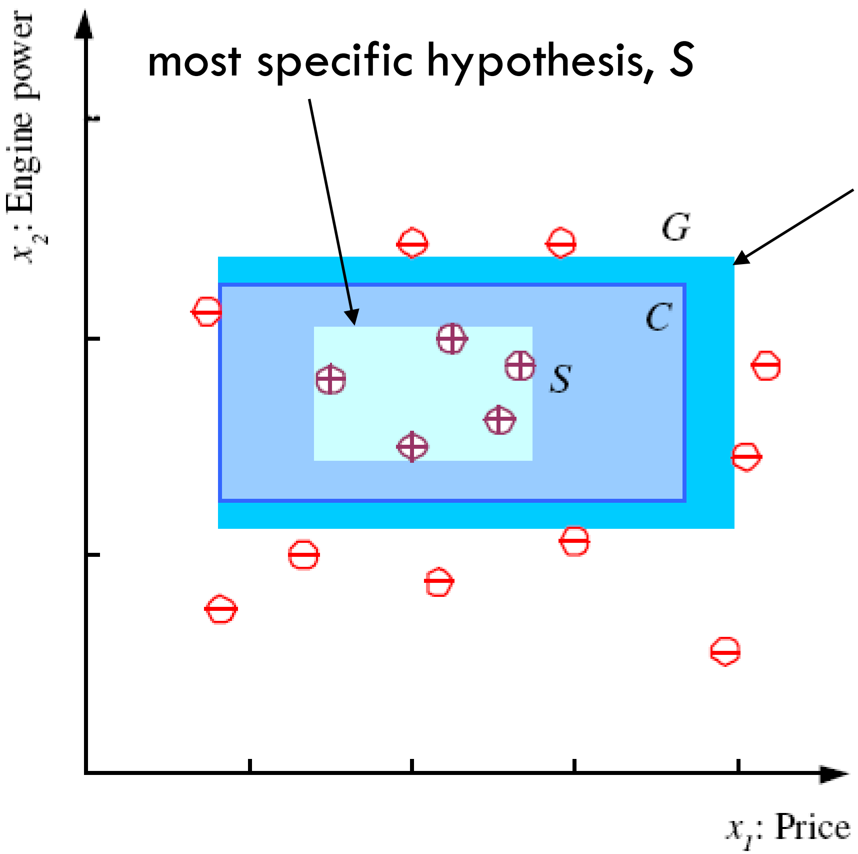


$$h(\mathbf{x}) = \begin{cases} 1 \text{ if } h \text{ says } \mathbf{x} \text{ is positive} \\ 0 \text{ if } h \text{ says } \mathbf{x} \text{ is negative} \end{cases}$$

Error of h on $\mathcal{H}$

$$E(h|\mathcal{X}) = \sum_{t=1}^{N} 1\left(h\left(\mathbf{x}^t\right) \neq r^t\right)$$

S, G, and the Version Space

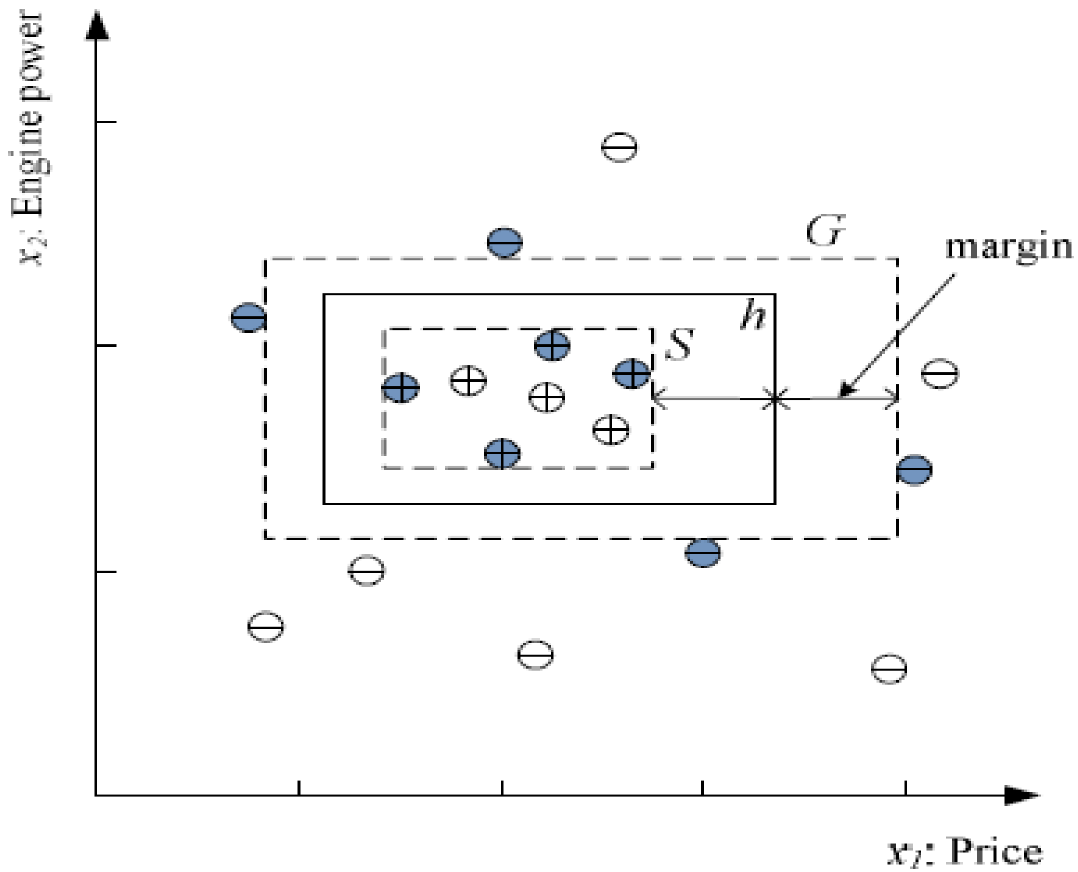


$h \in H$, between *S* and *G* is consistent and make up the version space

(Mitchell, 1997)

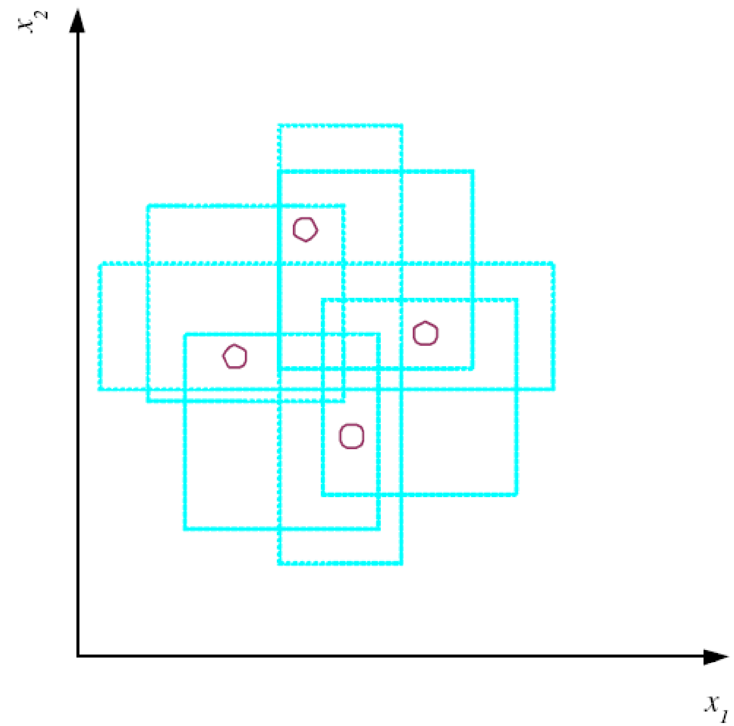
Margin

- Choose *h* with largest margin



VC Dimension

- N points can be labeled in 2^N ways as $+/-$
- $\mathcal{H}$ shatters N if there exists $h \in \mathcal{H}$ consistent for any of these: $VC(\mathcal{H}) = N$



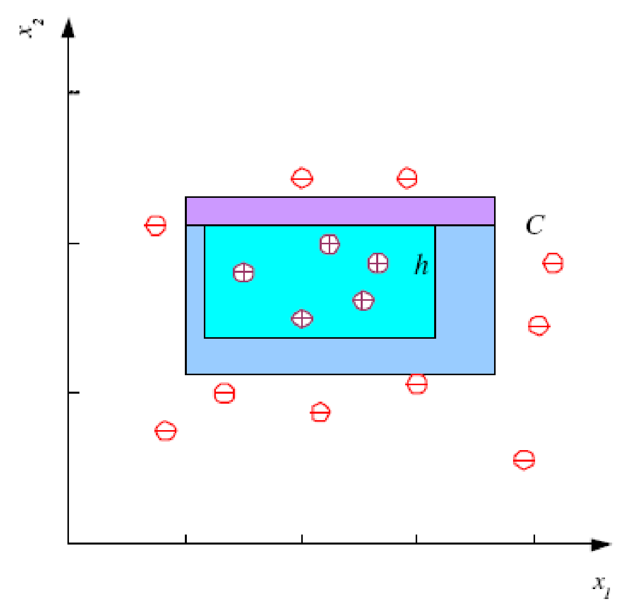
An axis-aligned rectangle shatters 4 points only !

Probably Approximately Correct (PAC) Learning

- How many training examples N should we have, such that with probability at least $1 - \delta$, h has error at most ε ?

  (Blumer et al., 1989)

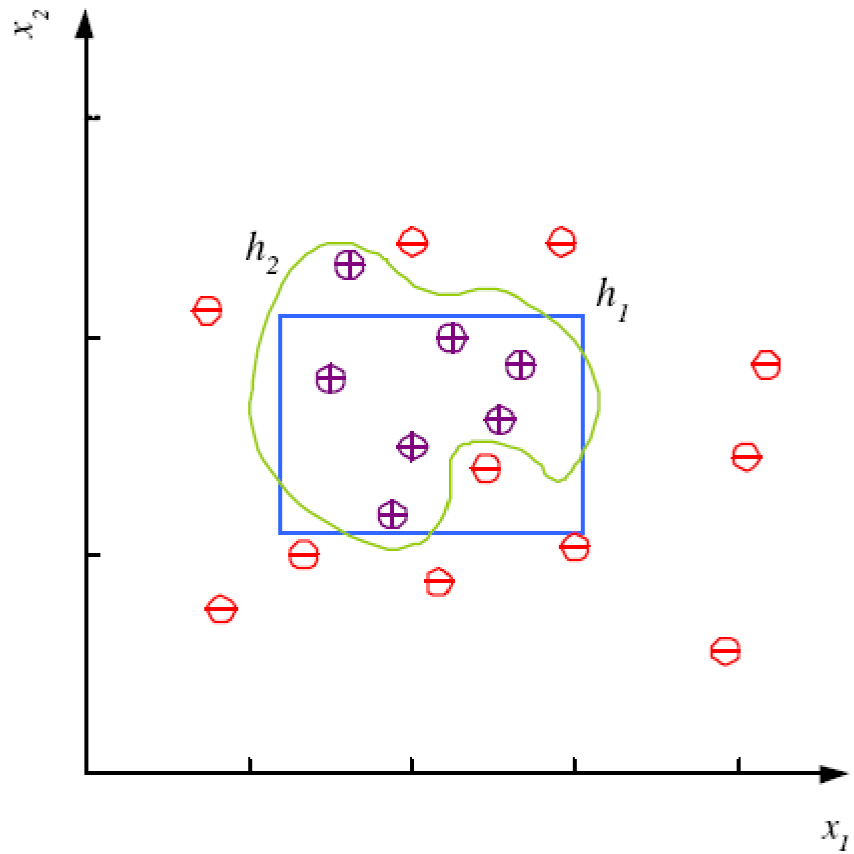
- Each strip is at most $\varepsilon/4$
- Pr that we miss a strip $1 - \varepsilon/4$
- Pr that N instances miss a strip $(1 - \varepsilon/4)^N$
- Pr that N instances miss 4 strips $4(1 - \varepsilon/4)^N$
- $4(1 - \varepsilon/4)^N \leq \delta$ and $(1 - x) \leq \exp(-x)$
- $4\exp(-\varepsilon N/4) \leq \delta$ and $N \geq (4/\varepsilon)\log(4/\delta)$



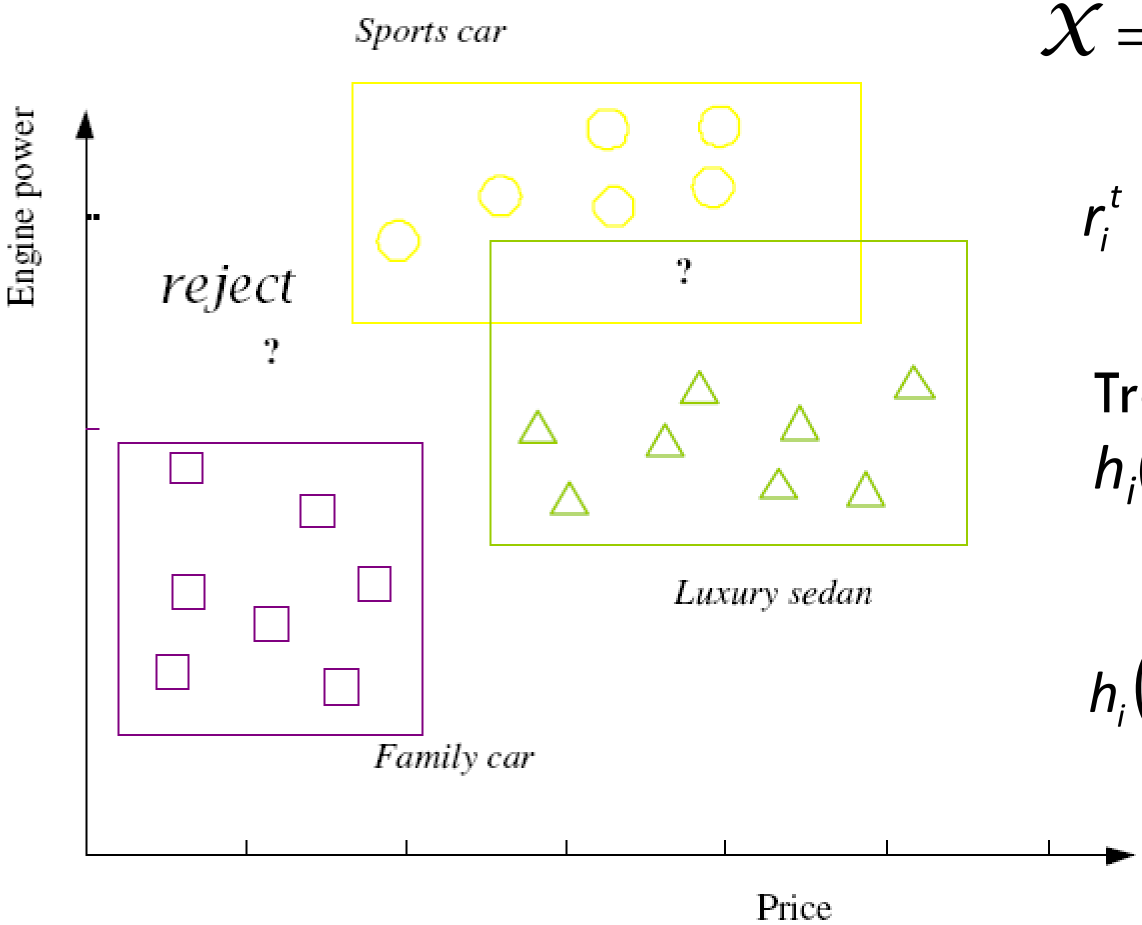
Noise and Model Complexity

Use the simpler one because

- Simpler to use (lower computational complexity)
- Easier to train (lower space complexity)
- Easier to explain (more interpretable)
- Generalizes better (lower variance - Occam's razor)



Multiple Classes, C_i i=1,...,K



$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^{N}$$

$$r_i^t = \begin{cases} 1 \text{ if } \mathbf{x}^t \in C_i \\ 0 \text{ if } \mathbf{x}^t \in C_j, j \neq i \end{cases}$$

Train hypotheses $h_i(\boldsymbol{x})$, $i = 1,...,K$:

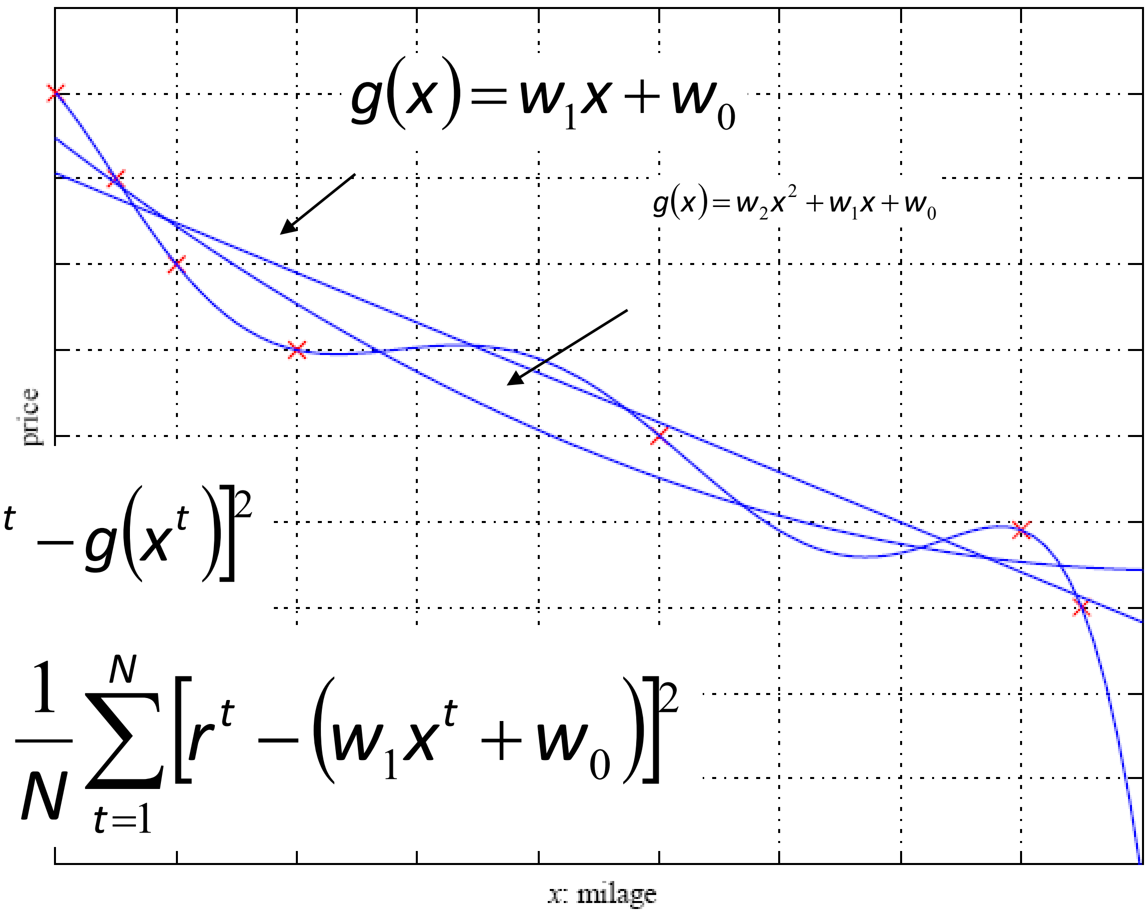
$$h_i(\mathbf{x}^t) = \begin{cases} 1 \text{ if } \mathbf{x}^t \in C_i \\ 0 \text{ if } \mathbf{x}^t \in C_j, j \neq i \end{cases}$$

Regression

$$\mathcal{X} = \{x^t, r^t\}_{t=1}^N$$

$$r^t \in \Re$$

$$r^t = f(x^t) + \varepsilon$$



$$E(g \mid \mathcal{X}) = \frac{1}{N} \sum_{t=1}^{N} \left[r^t - g(x^t) \right]^2$$

$$E(w_1, w_0 \mid \mathcal{X}) = \frac{1}{N} \sum_{t=1}^{N} \left[r^t - (w_1 x^t + w_0) \right]^2$$

Model Selection & Generalization

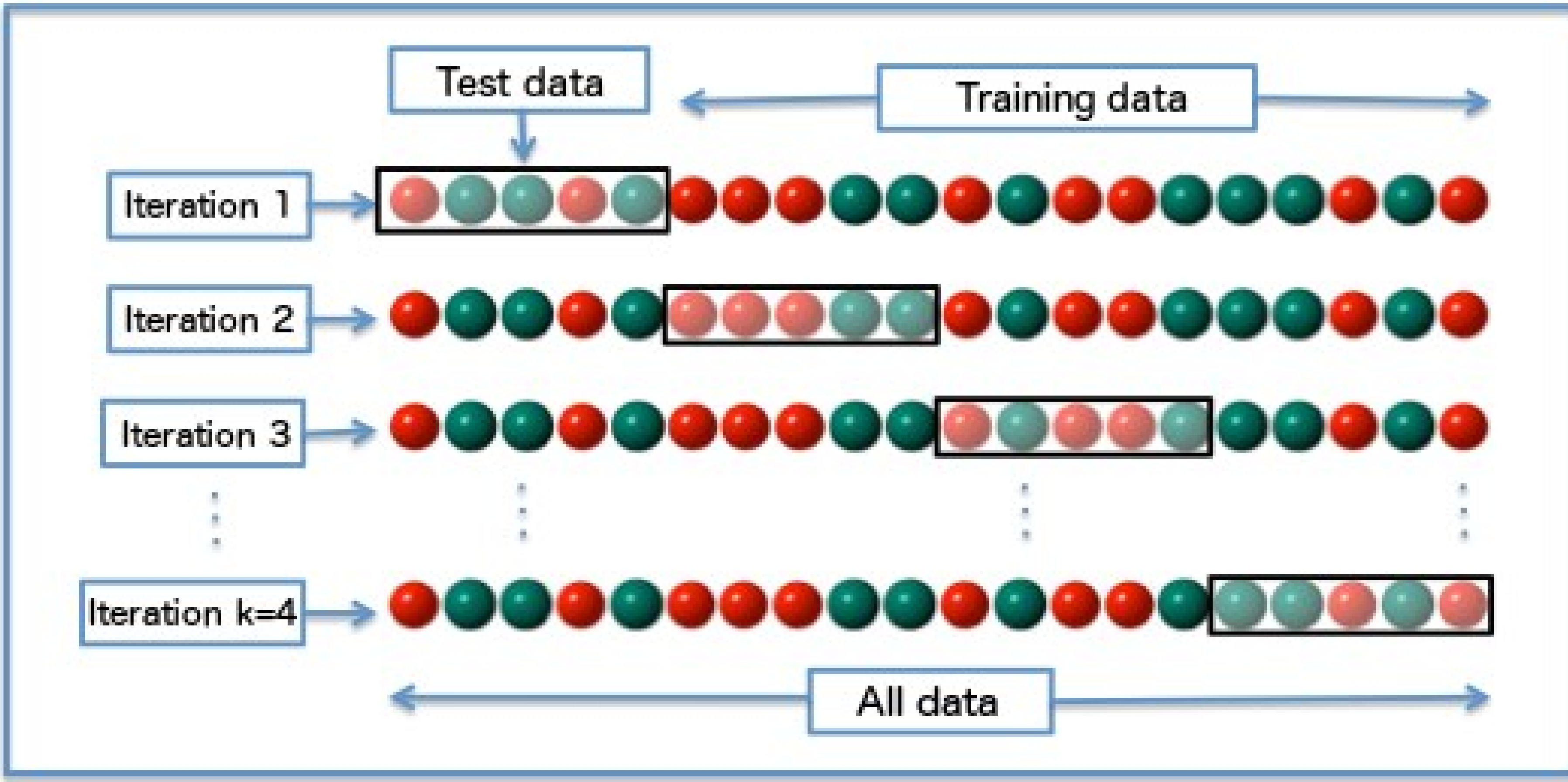
- Learning is an ill-posed problem; data is not sufficient to find a unique solution
- The need for inductive bias, assumptions about $\mathcal{H}$
- Generalization: How well a model performs on new data
- Overfitting: $\mathcal{H}$ more complex than C or f
- Underfitting: $\mathcal{H}$ less complex than C or f

Triple Trade-Off

- There is a trade-off between three factors (Dietterich, 2003):
  1. Complexity of $\mathcal{H}$, c ($\mathcal{H}$),
  2. Training set size, N,
  3. Generalization error, E, on new data
- As N, $E\downarrow$
- As c ($\mathcal{H}$), first $E\downarrow$ and then E

Cross-Validation

- To estimate generalization error, we need data unseen during training. We split the data as
  - Training set (50%)
  - Validation set (25%)
  - Test (publication) set (25%)
- Resampling when there is few data


Test data
Training data
Iteration 1
Iteration 2
Iteration 3
Iteration k=4
All data

Dimensions of a Supervised Learner

1. Model: $g(\mathbf{x} \mid \theta)$

2. Loss function: $E(\theta \mid \mathcal{X}) = \sum_t L\left(r^t, g\left(\mathbf{x}^t \mid \theta\right)\right)$

3. Optimization procedure:

$$\theta^* = \arg\min_\theta E(\theta \mid \mathcal{X})$$

CHAPTER 3:

BAYESİAN DECİSİON THEORY

NAIVE BAYESIAN LEARNING/CLASSIFICATION

- Simple probabilistic classifiers
- Naïve: Strong independence assumptions between the features
- From 1950s onwards
- Text mining: Word frequencies as features
- Maximum likelihood (training)
- For too many features (n=10.000) or a feature obtaining a large value: infeasible
- Apple example: Supervised learning setting

Is this fruit an apple?

- Features (contribute independently):
  - Red
  - Round
  - 10 cm in diameter
- Conclusion → Apple?

Probabilistic model

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

WHAT WOULD A SUITABLE EXAMPLE BE?

Prior Probability

Likelihood of the evidence 'E' if the Hypothesis 'H' is true

$$P(H|E) = \frac{P(H) * P(E|H)}{P(E)}$$

Posterior Probability of 'H' given the evidence

Priori probability that the evidence itself is true

Likelihood

How probable is the evidence given that our hypothesis is true?

Prior

How probable was our hypothesis *before* observing the evidence?

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

Posterior

How probable is our hypothesis given the observed evidence?
(Not directly computable)

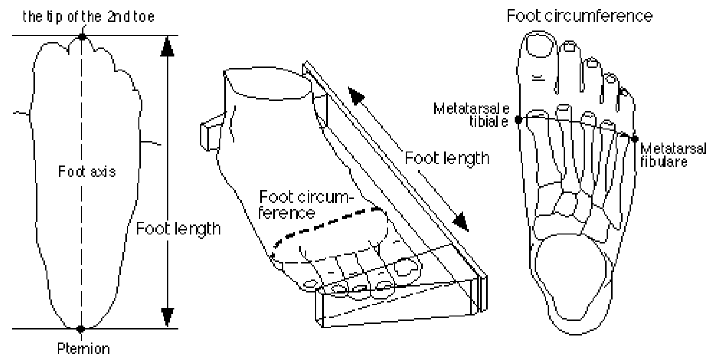
Marginal

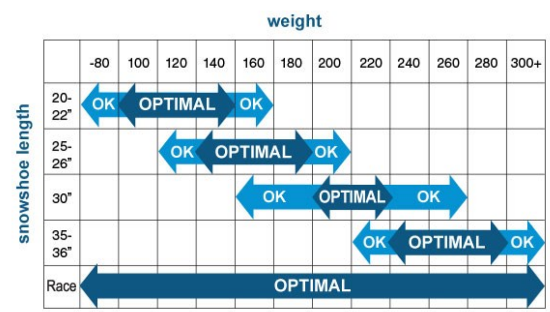
How probable is the new evidence under all possible hypotheses?

$P(e) = \sum P(e \mid H_i)\, P(H_i)$

Gender classification / Training set

Gender	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9





Wikipedia,
2017

MEANS & VARIANCE → TESTING

Gender	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

Gender	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

GAUSSIAN NAIVE BAYES

$$p(x = v \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}}\, e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

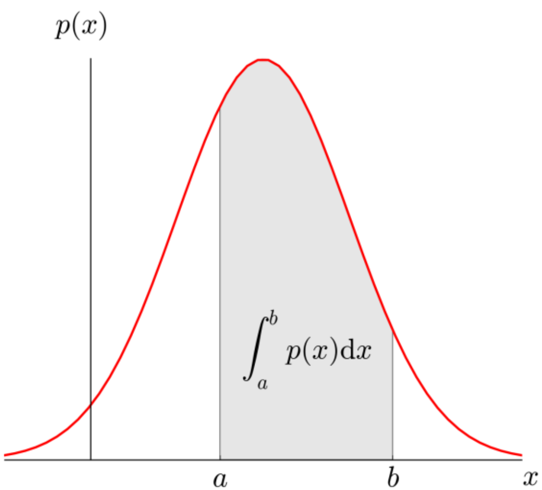
$$\text{posterior (male)} = \frac{P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male})}{evidence}$$

$$\text{posterior (female)} = \frac{P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})}{evidence}$$

AN EXAMPLE FOR THE CALCULATION

$$P(\text{male}) = 0.5$$

$$p(\text{height} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789,$$



MALE OR FEMALE?

$$p(\text{weight} \mid \text{male}) = 5.9881 \cdot 10^{-6}$$
$$p(\text{foot size} \mid \text{male}) = 1.3112 \cdot 10^{-3}$$
$$\text{posterior numerator (male)} = \text{their product} = 6.1984 \cdot 10^{-9}$$

$$P(\text{female}) = 0.5$$
$$p(\text{height} \mid \text{female}) = 2.2346 \cdot 10^{-1}$$
$$p(\text{weight} \mid \text{female}) = 1.6789 \cdot 10^{-2}$$
$$p(\text{foot size} \mid \text{female}) = 2.8669 \cdot 10^{-1}$$
$$\text{posterior numerator (female)} = \text{their product} = 5.3778 \cdot 10^{-4}$$

Probability and Inference

- Result of tossing a coin is $\in$ {Heads,Tails}
- Random var $X \in \{1,0\}$

  Bernoulli: $P\{X=1\} = p_o^{X}(1 - p_o)^{(1 - X)}$
- Sample: $\mathbf{X} = \{x^t\}_{t=1}^{N}$

  Estimation: $p_o = \#\{\text{Heads}\}/\#\{\text{Tosses}\} = \sum_t x^t / N$
- Prediction of next toss:

  Heads if $p_o > ½$, Tails otherwise

Classification

- Credit scoring: Inputs are income and savings.
  Output is low-risk vs high-risk
- Input: $\mathbf{x} = [x_1, x_2]^T$,Output: C Î {0,1}
- Prediction:

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C=1 \mid x_1, x_2) > 0.5 \\ C = 0 \text{ otherwise} \end{cases}$$

or

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C=1 \mid x_1, x_2) > P(C=0 \mid x_1, x_2) \\ C = 0 \text{ otherwise} \end{cases}$$

Bayes' Rule

prior *likelihood*

posterior

$$P(C \mid \mathbf{x}) = \frac{P(C)\,p(\mathbf{x} \mid C)}{p(\mathbf{x})}$$

evidence

$$P(C=0) + P(C=1) = 1$$

$$p(\mathbf{x}) = p(\mathbf{x} \mid C=1)P(C=1) + p(\mathbf{x} \mid C=0)P(C=0)$$

$$p(C=0 \mid \mathbf{x}) + P(C=1 \mid \mathbf{x}) = 1$$

Bayes' Rule: K>2 Classes

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

$$= \frac{p(\mathbf{x} \mid C_i)P(C_i)}{\sum_{k=1}^{K} p(\mathbf{x} \mid C_k)P(C_k)}$$

$P(C_i) \geq 0$ and $\sum_{i=1}^{K} P(C_i) = 1$

choose C_i if $P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$

Losses and Risks

- Actions: α_i
- Loss of α_i when the state is C_k : λ_{ik}
- Expected risk (Duda and Hart, 1973)

$$R(\alpha_i \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \mathbf{x})$$

$$\text{choose } \alpha_i \text{ if } R(\alpha_i \mid \mathbf{x}) = \min_k R(\alpha_k \mid \mathbf{x})$$

Losses and Risks: 0/1 Loss

$$\lambda_{ik} = \begin{cases} 0 \text{ if } i = k \\ 1 \text{ if } i \neq k \end{cases}$$

$$\begin{aligned} R(\alpha_i \mid \mathbf{x}) &= \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \mathbf{x}) \\ &= \sum_{k \neq i} P(C_k \mid \mathbf{x}) \\ &= 1 - P(C_i \mid \mathbf{x}) \end{aligned}$$

For minimum risk, choose the most probable class

Losses and Risks: Reject

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ \lambda & \text{if } i = K+1, \quad 0 < \lambda < 1 \\ 1 & \text{otherwise} \end{cases}$$

$$R(\alpha_{K+1} \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda P(C_k \mid \mathbf{x}) = \lambda$$

$$R(\alpha_i \mid \mathbf{x}) = \sum_{k \neq i} P(C_k \mid \mathbf{x}) = 1 - P(C_i \mid \mathbf{x})$$

choose C_i if $P(C_i \mid \mathbf{x}) > P(C_k \mid \mathbf{x}) \ \forall k \neq i$ and $P(C_i \mid \mathbf{x}) > 1 - \lambda$

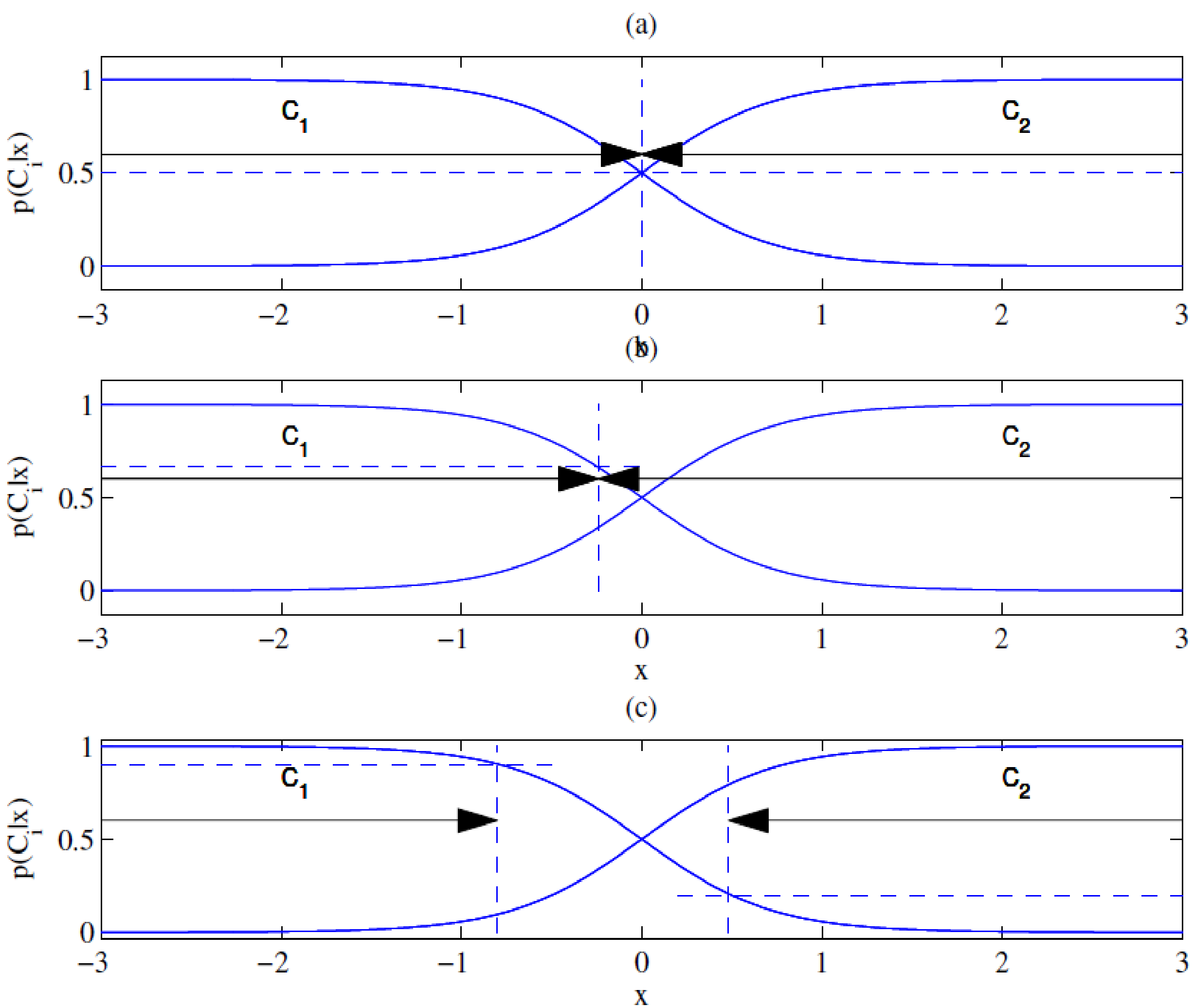
reject otherwise

Different Losses and Reject

Equal losses

Unequal losses

With reject

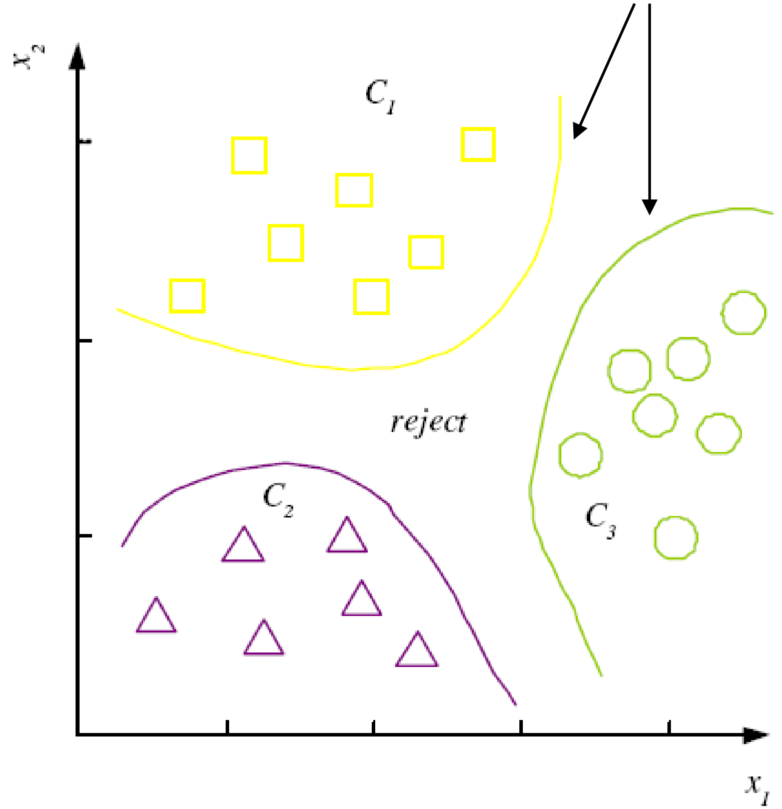


Discriminant Functions

choose C_i if $g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})$

$$g_i(\mathbf{x}), i = 1, \ldots, K$$

$$g_i(\mathbf{x}) = \begin{cases} -R(\alpha_i \mid \mathbf{x}) \\ P(C_i \mid \mathbf{x}) \\ p(\mathbf{x} \mid C_i)P(C_i) \end{cases}$$



K decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_K$

$$\mathcal{R}_i = \{\mathbf{x} \mid g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$

K=2 Classes

- Dichotomizer (K=2) vs Polychotomizer (K>2)
- $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$

$$\text{choose} \begin{cases} C_1 \text{ if } g(\mathbf{x}) > 0 \\ C_2 \text{ otherwise} \end{cases}$$

- *Log odds:* $\log \dfrac{P(C_1 \mid \mathbf{x})}{P(C_2 \mid \mathbf{x})}$

Utility Theory

- Prob of state k given exidence $\mathbf{x}$: $P(S_k \mid \mathbf{x})$
- Utility of α_i when state is k: U_{ik}
- Expected utility:

$$EU(\alpha_i \mid \mathbf{x}) = \sum_k U_{ik} P(S_k \mid \mathbf{x})$$

$$\text{Choose } \alpha_i \text{ if } EU(\alpha_i \mid \mathbf{x}) = \max_j EU(\alpha_j \mid \mathbf{x})$$

Association Rules

- Association rule: $X \rightarrow Y$
- *People who buy/click/visit/enjoy X are also likely to buy/click/visit/enjoy Y.*
- A rule implies association, not necessarily causation.

Association measures

- Support ($X \rightarrow Y$):

$$P(X,Y) = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers}\}}$$

- Confidence ($X \rightarrow Y$):

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)} = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

- Lift ($X \rightarrow Y$):

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

Example

Transaction	Items in basket
1	milk, bananas, chocolate
2	milk, chocolate
3	milk, bananas
4	chocolate
5	chocolate
6	milk, chocolate

SOLUTION:

$$\begin{aligned}
\text{milk} \rightarrow \text{bananas} &\;:\; \text{Support} = 2/6, \text{Confidence} = 2/4 \\
\text{bananas} \rightarrow \text{milk} &\;:\; \text{Support} = 2/6, \text{Confidence} = 2/2 \\
\text{milk} \rightarrow \text{chocolate} &\;:\; \text{Support} = 3/6, \text{Confidence} = 3/4 \\
\text{chocolate} \rightarrow \text{milk} &\;:\; \text{Support} = 3/6, \text{Confidence} = 3/5
\end{aligned}$$

Apriori algorithm (Agrawal et al., 1996)

- For (X,Y,Z), a 3-item set, to be frequent (have enough support), (X,Y), (X,Z), and (Y,Z) should be frequent.
- If (X,Y) is not frequent, none of its supersets can be frequent.
- Once we find the frequent *k*-item sets, we convert them to rules: X, Y $\rightarrow$ Z, ...

  and X $\rightarrow$ Y, Z, ...

CHAPTER 4:

PARAMETRİC METHODS

Parametric Estimation

- $\mathcal{X} = \{ x^t \}_t$ where $x^t \sim p(x)$
- Parametric estimation:

  Assume a form for $p(x \mid \theta)$ and estimate θ, its sufficient statistics, using X

  e.g., N (μ, σ^2) where $\theta = \{ \mu, \sigma^2 \}$

Maximum Likelihood Estimation

- Likelihood of θ given the sample $\mathcal{X}$

$$l\,(\vartheta\,|\,\mathcal{X}) = p\,(\mathcal{X}\,|\,\vartheta) = \prod_t p\,(x^t\,|\,\vartheta)$$

- Log likelihood

$$\mathcal{L}(\vartheta\,|\,\mathcal{X}) = \log l\,(\vartheta\,|\,\mathcal{X}) = \sum_t \log p\,(x^t\,|\,\vartheta)$$

- Maximum likelihood estimator (MLE)

$$\vartheta^* = \operatorname{argmax}_\vartheta \mathcal{L}(\vartheta\,|\,\mathcal{X})$$

Examples: Bernoulli/Multinomial

- Bernoulli: Two states, failure/success, x in {0,1}

  $P(x) = p_o^{x} (1 - p_o)^{(1 - x)}$

  $\mathcal{L}(p_o | \mathcal{X}) = \log \prod_t p_o^{x^t} (1 - p_o)^{(1 - x^t)}$

  MLE: $p_o = \sum_t x^t / N$

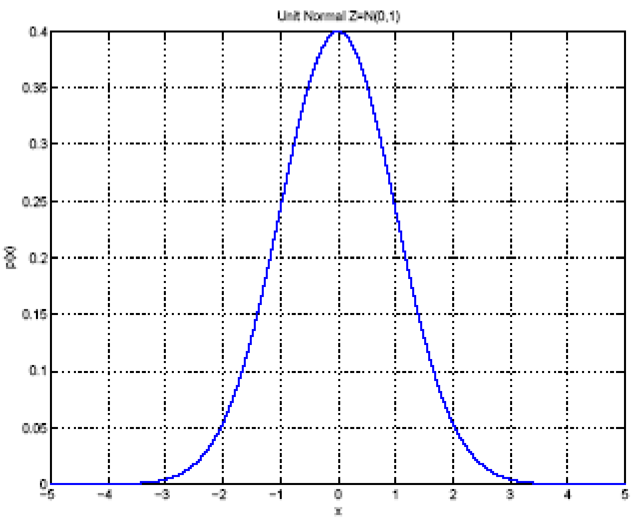
- Multinomial: $K>2$ states, x_i in {0,1}

  $P(x_1, x_2, \ldots, x_K) = \prod_i p_i^{x_i}$

  $\mathcal{L}(p_1, p_2, \ldots, p_K | \mathcal{X}) = \log \prod_t \prod_i p_i^{x_i^t}$

  MLE: $p_i = \sum_t x_i^t / N$

Gaussian (Normal) Distribution



$$m = \frac{\sum_t x^t}{N}$$

$$s^2 = \frac{\sum_t \left(x^t - m\right)^2}{N}$$

- $p(x) = \mathcal{N}(\mu, \sigma^2)$

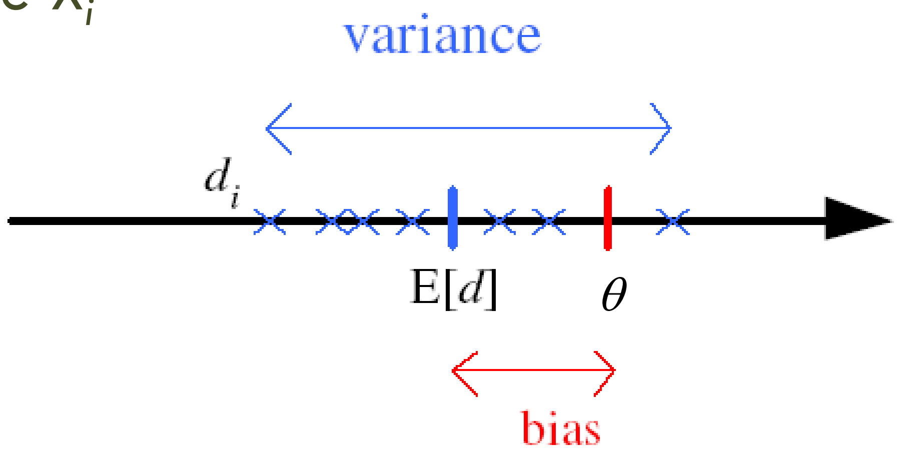
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- MLE for μ and σ^2:

Bias and Variance

Unknown parameter θ

Estimator $d_i = d\,(X_i)$ on sample X_i



Bias: $b_\theta(d) = E\,[d] - \theta$

Variance: $E\,[(d - E\,[d])^2]$

Mean square error:

$$r\,(d,\theta) = E\,[(d-\theta)^2]$$
$$= (E\,[d] - \theta)^2 + E\,[(d-E\,[d])^2]$$
$$= \text{Bias}^2 + \text{Variance}$$

Bayes' Estimator

- Treat ϑ as a random var with prior $p\ (\vartheta)$
- Bayes' rule: $p\ (\vartheta|\mathcal{X}) = p(\mathcal{X}|\vartheta)\ p(\vartheta)\ /\ p(\mathcal{X})$
- Full: $p(x|\mathcal{X}) = \int p(x|\vartheta)\ p(\vartheta|\mathcal{X})\ d\vartheta$
- Maximum a Posteriori (MAP):

  $$\vartheta_{MAP} = \mathrm{argmax}_{\vartheta}\ p(\vartheta|\mathcal{X})$$
- Maximum Likelihood (ML): $\vartheta_{ML} = \mathrm{argmax}_{\vartheta}\ p(\mathcal{X}|\vartheta)$
- Bayes': $\vartheta_{Bayes'} = E[\vartheta|\mathcal{X}] = \int \vartheta\ p(\vartheta|\mathcal{X})\ d\vartheta$

Bayes' Estimator: Example

- $x^t \sim \mathcal{N}(\vartheta, \sigma_o^2)$ and $\vartheta \sim \mathcal{N}(\mu, \sigma^2)$
- $\vartheta_{ML} = m$
- $\vartheta_{MAP} = \vartheta_{Bayes'} =$

$$E[\theta \mid \mathcal{X}] = \frac{N/\sigma_0^2}{N/\sigma_0^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_0^2 + 1/\sigma^2} \mu$$

Parametric Classification

$$g_i(x) = p(x \mid C_i)P(C_i)$$

or

$$g_i(x) = \log p(x \mid C_i) + \log P(C_i)$$

$$p(x \mid C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

- Given the sample $\mathcal{X} = \{x^t, r^t\}_{t=1}^N$

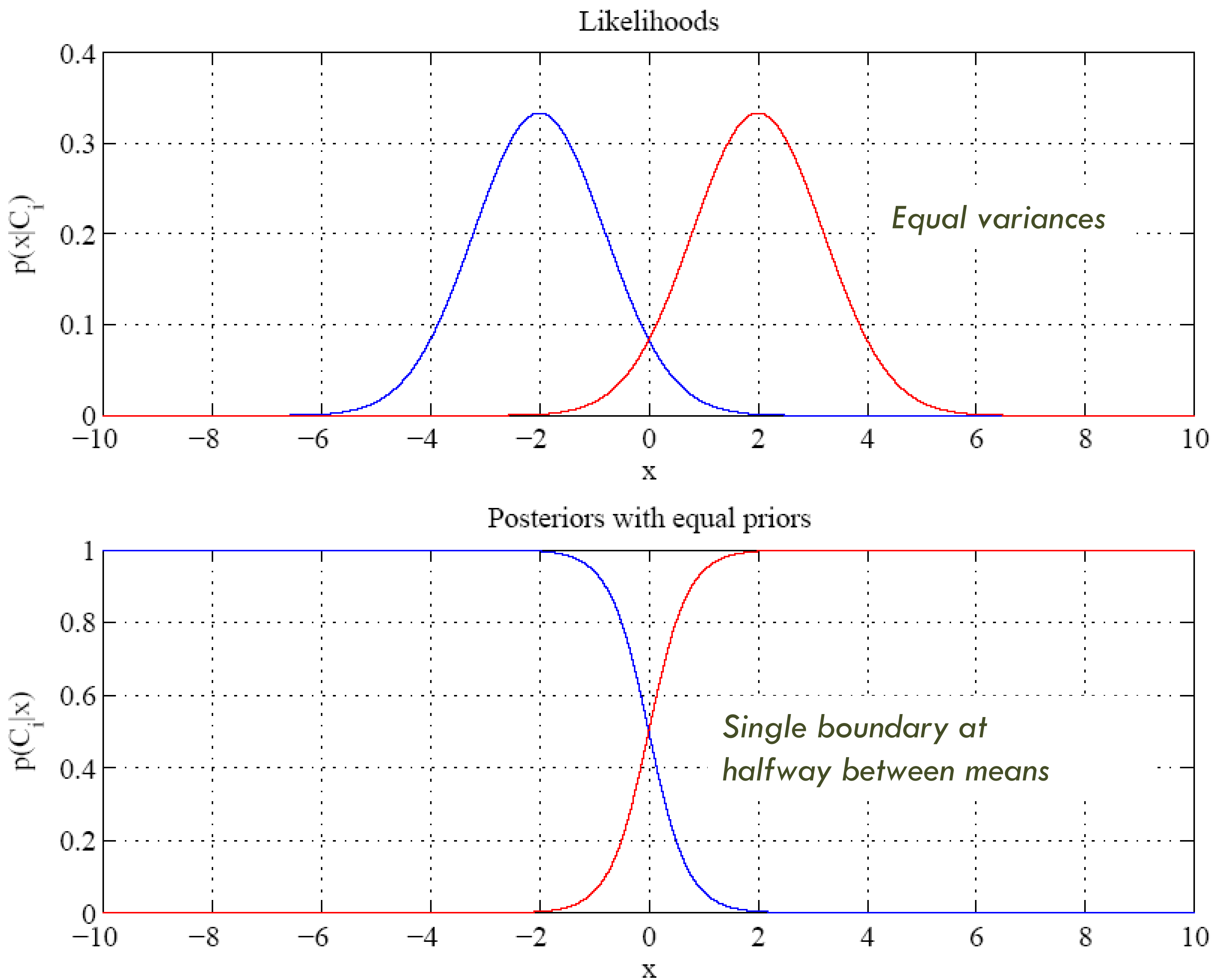
$$x \in \Re \qquad r_i^t = \begin{cases} 1 \text{ if } x^t \in C_i \\ 0 \text{ if } x^t \in C_j, j \neq i \end{cases}$$

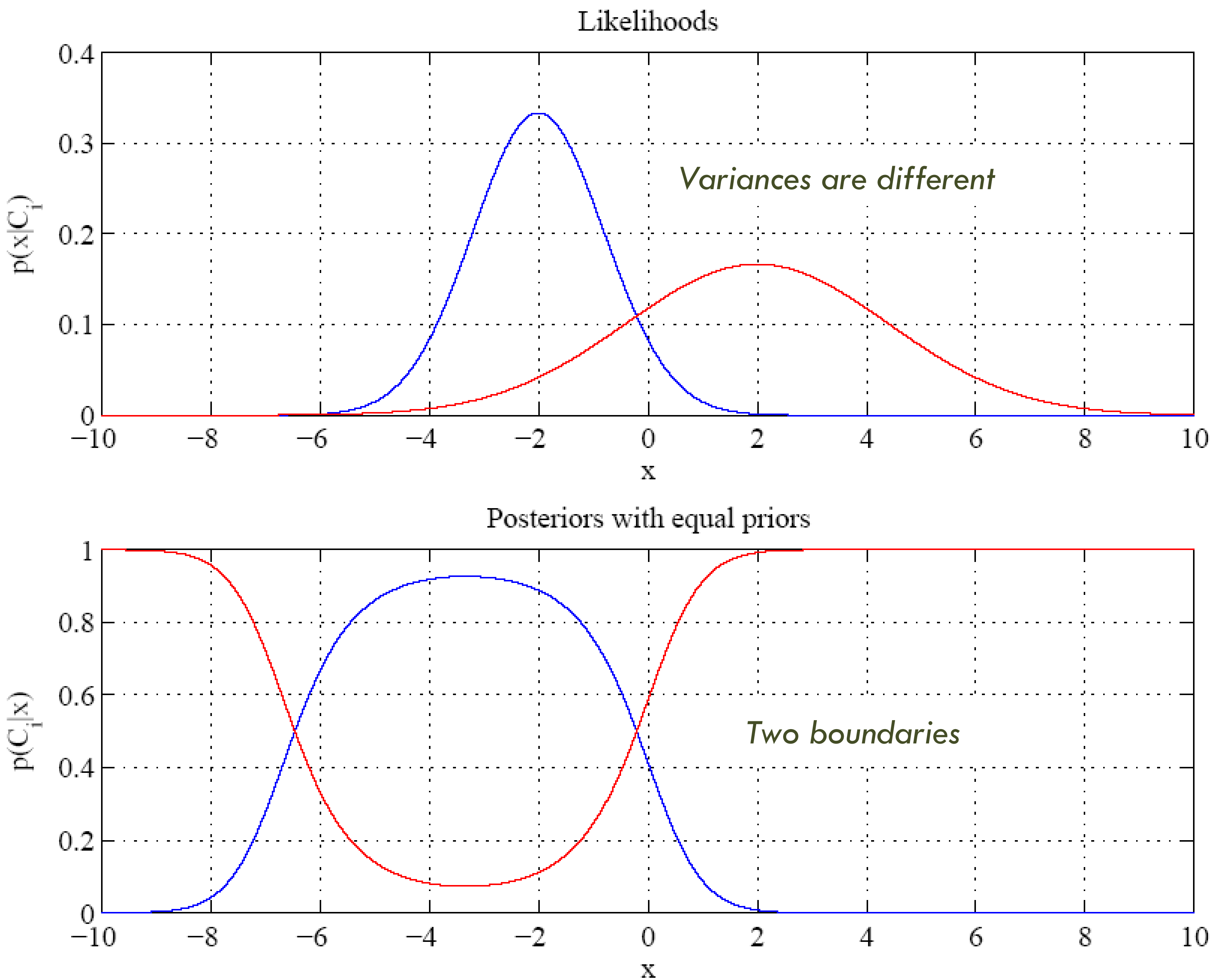
- ML estimates are

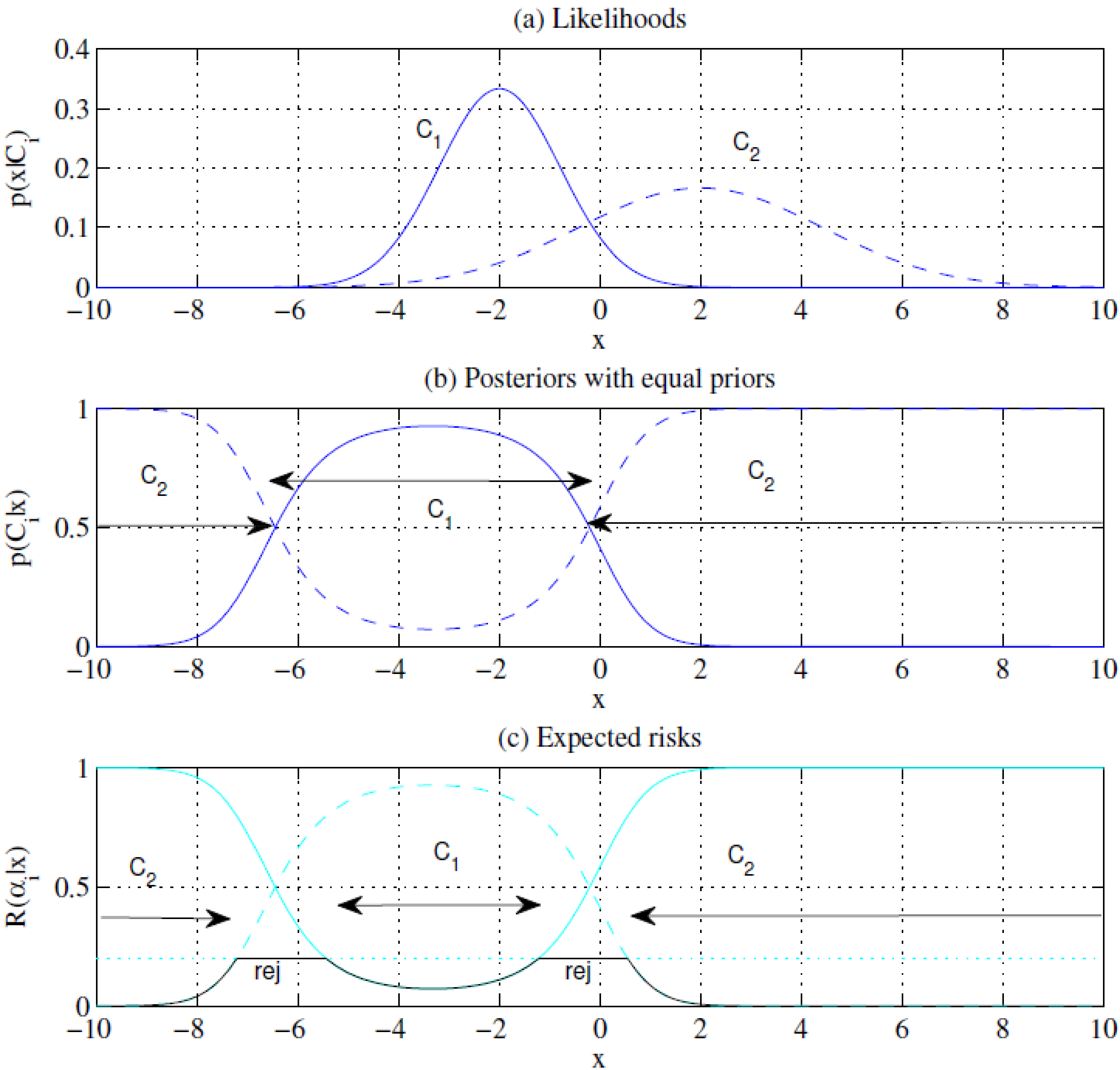
$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \quad s_i^2 = \frac{\sum_t \left(x^t - m_i\right)^2 r_i^t}{\sum_t r_i^t}$$

- Discriminant

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{\left(x - m_i\right)^2}{2 s_i^{\,2}} + \log \hat{P}(C_i)$$


Likelihoods
p(x|C_i)
x
Equal variances
Posteriors with equal priors
p(C_i|x)
x
Single boundary at
halfway between means


Likelihoods
Variances are different
p(x|C_i)
0.4
0.3
0.2
0.1
0
−10
−8
−6
−4
−2
0
2
4
6
8
10
x
Posteriors with equal priors
Two boundaries
p(C_i|x)
1
0.8
0.6
0.4
0.2
0
−10
−8
−6
−4
−2
0
2
4
6
8
10
x


(a) Likelihoods
p(x|C_i)
C_1
C_2
x
(b) Posteriors with equal priors
p(C_i|x)
C_2
C_1
C_2
x
(c) Expected risks
R(α_i|x)
C_2
C_1
C_2
rej
rej
x

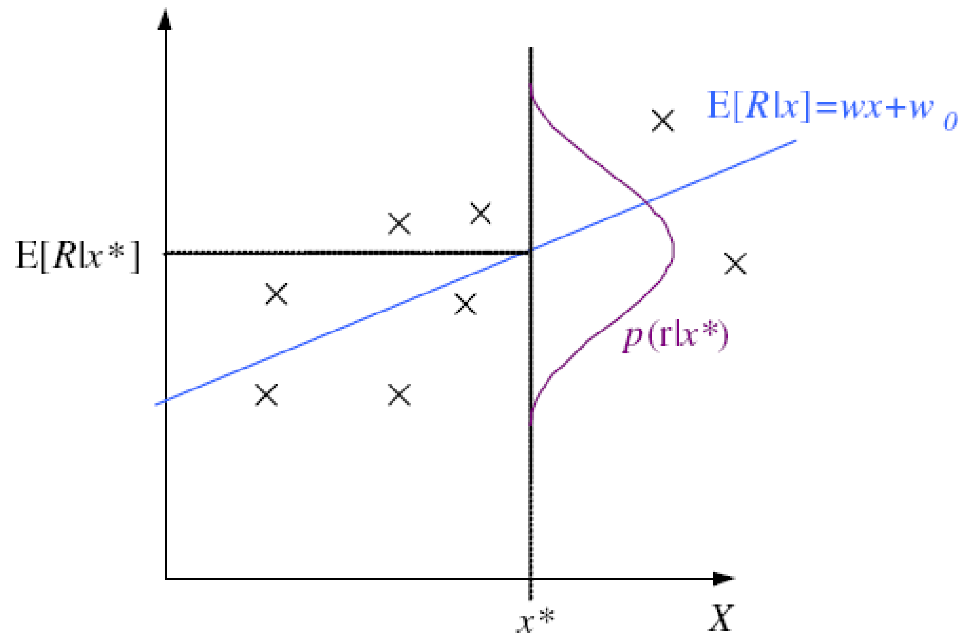
Regression

$$r = f(x) + \varepsilon$$

$$\text{estimator}: g(x \mid \theta)$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$p(r \mid x) \sim \mathcal{N}(g(x \mid \theta), \sigma^2)$$



$$\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{t=1}^{N} p(x^t, r^t)$$

$$= \log \prod_{t=1}^{N} p(r^t \mid x^t) + \log \prod_{t=1}^{N} p(x^t)$$

Regression: From LogL to Error

$$\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left[r^t - g\left(x^t \mid \theta\right)\right]^2}{2\sigma^2}\right]$$

$$= -N \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{t=1}^{N} \left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

$$E(\theta \mid \mathcal{X}) = \frac{1}{2} \sum_{t=1}^{N} \left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

Linear Regression

$$g\left(x^t \mid w_1, w_0\right) = w_1 x^t + w_0$$

$$\sum_t r^t = N w_0 + w_1 \sum_t x^t$$

$$\sum_t r^t x^t = w_0 \sum_t x^t + w_1 \sum_t \left(x^t\right)^2$$

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t \left(x^t\right)^2 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \end{bmatrix}$$

$$\mathbf{w} = \mathbf{A}^{-1} \mathbf{y}$$

Polynomial Regression

$$g\left(x^t \mid w_k, \ldots, w_2, w_1, w_0\right) = w_k\left(x^t\right)^k + \cdots + w_2\left(x^t\right)^2 + w_1 x^t + w_0$$

$$\mathbf{D} = \begin{bmatrix} 1 & x^1 & \left(x^1\right)^2 & \cdots & \left(x^1\right)^k \\ 1 & x^2 & \left(x^2\right)^2 & \cdots & \left(x^2\right)^k \\ \vdots & & & & \\ 1 & x^N & \left(x^N\right)^2 & \cdots & \left(x^N\right)^2 \end{bmatrix} \quad \mathbf{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}$$

$$\mathbf{w} = \left(\mathbf{D}^T \mathbf{D}\right)^{-1} \mathbf{D}^T \mathbf{r}$$

Other Error Measures

- Square Error: $E(\theta|\mathcal{X})=\frac{1}{2}\sum_{t=1}^{N}\left[r^t-g\left(x^t|\theta\right)\right]^2$
- Relative Square Error: $$E(\theta|\mathcal{X})=\frac{\sum_{t=1}^{N}\left[r^t-g\left(x^t|\theta\right)\right]^2}{\sum_{t=1}^{N}\left[r^t-\bar{r}\right]^2}$$
- Absolute Error: $E(\vartheta|X)=\sum_t |r^t-g(x^t|\vartheta)|$
- ε-sensitive Error:

$$E(\vartheta|X)=\sum_t 1(|r^t-g(x^t|\vartheta)|>\varepsilon)\,(|r^t-g(x^t|\vartheta)|-\varepsilon)$$

Bias and Variance

$$E\left[(r-g(x))^2 \mid x\right] = E\left[(r-E[r \mid x])^2 \mid x\right] + (E[r \mid x] - g(x))^2$$

noise *squared error*

$$E_X\left[(E[r \mid x] - g(x))^2 \mid x\right] = (E[r \mid x] - E_X[g(x)])^2 + E_X\left[(g(x) - E_X[g(x)])^2\right]$$

bias *variance*

Estimating Bias and Variance

- M samples $X_i=\{x^t_i, r^t_i\}$, $i=1,\dots,M$ are used to fit $g_i(x)$, $i=1,\dots,M$

$$\text{Bias}^2(g)=\frac{1}{N}\sum_t\left[\bar{g}\left(x^t\right)-f\left(x^t\right)\right]^2$$

$$\text{Variance}(g)=\frac{1}{NM}\sum_t\sum_i\left[g_i\left(x^t\right)-\bar{g}\left(x^t\right)\right]^2$$

$$\bar{g}(x)=\frac{1}{M}\sum_t g_i(x)$$

Bias/Variance Dilemma

- Example: $g_i(x)=2$ has no variance and high bias

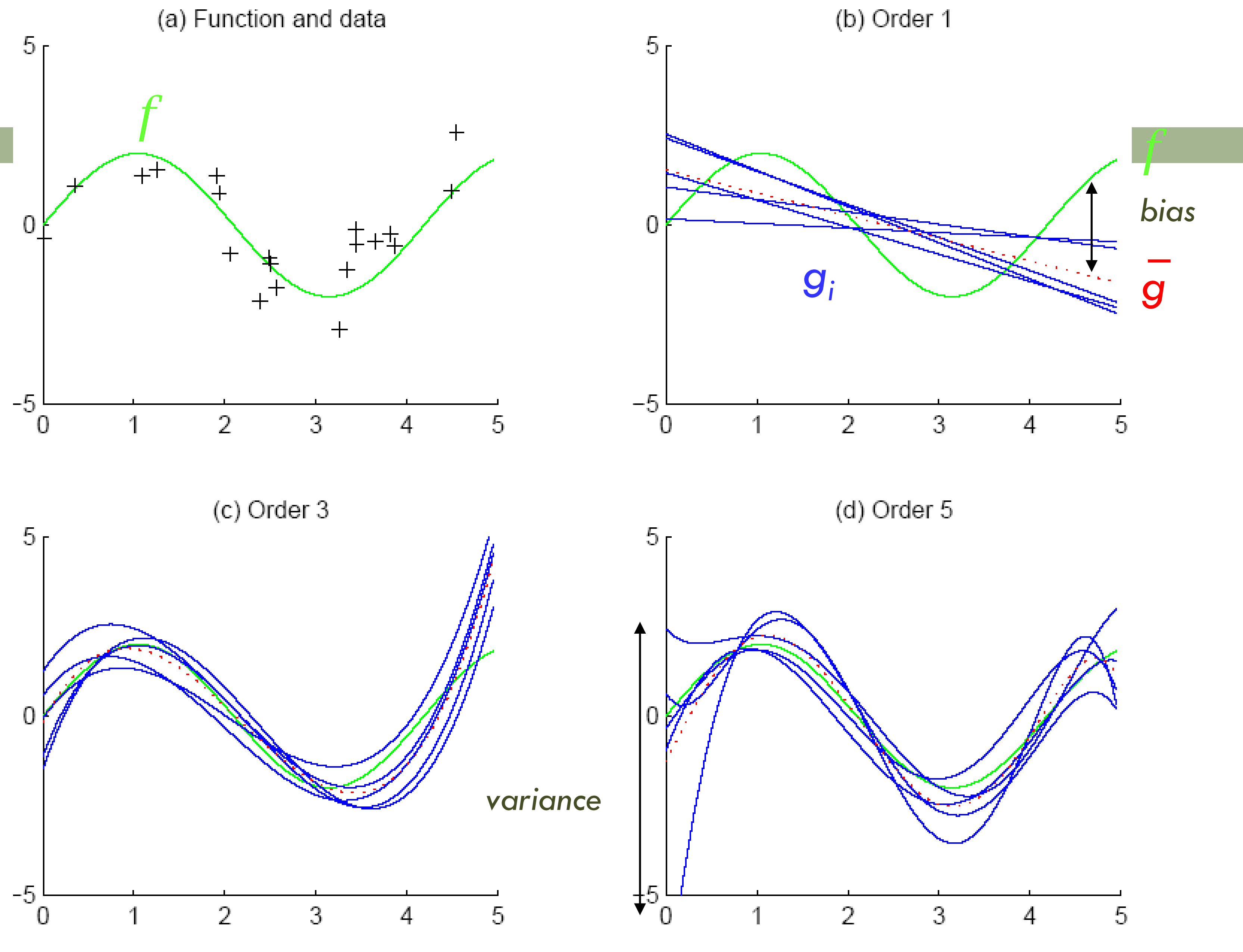
  $g_i(x)= \sum_t r^t_i/N$ has lower bias with variance

- As we increase complexity,

  bias decreases (a better fit to data) and

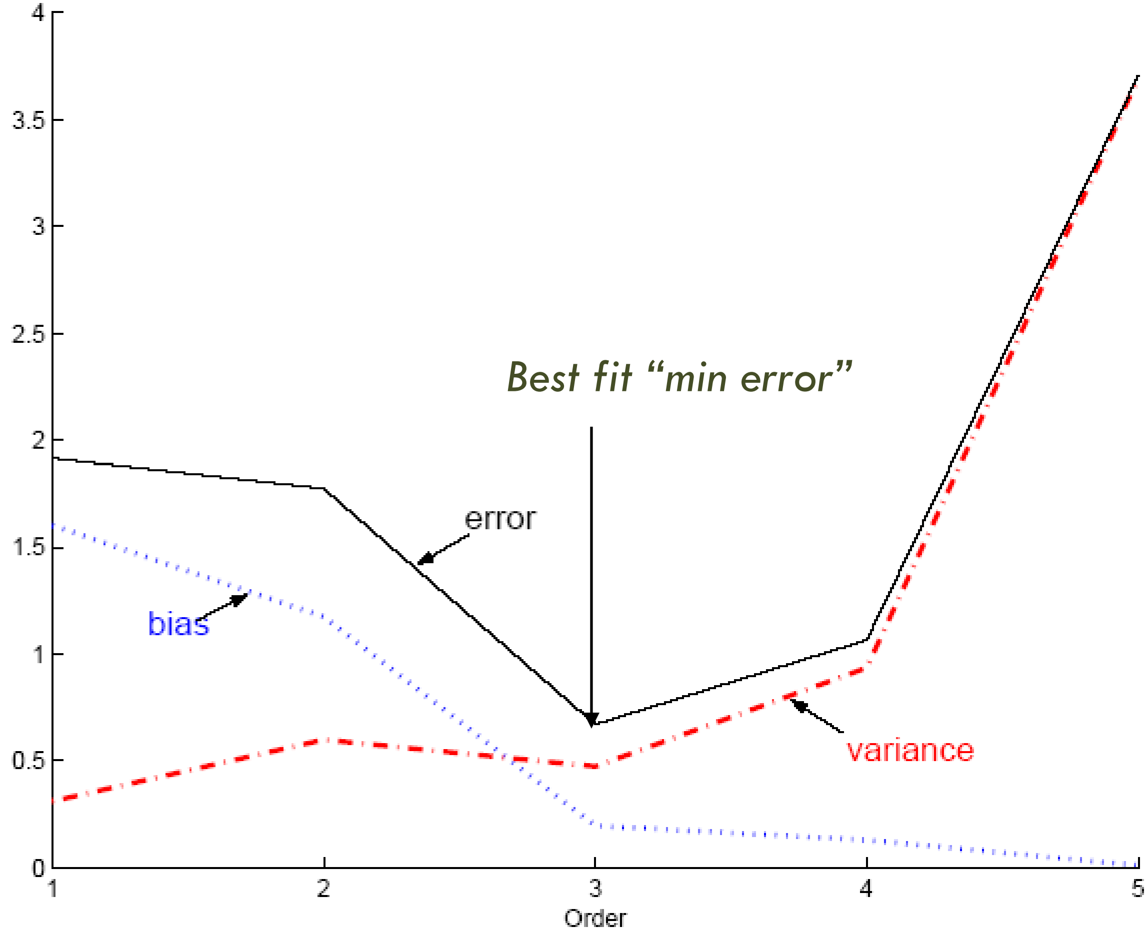
  variance increases (fit varies more with data)

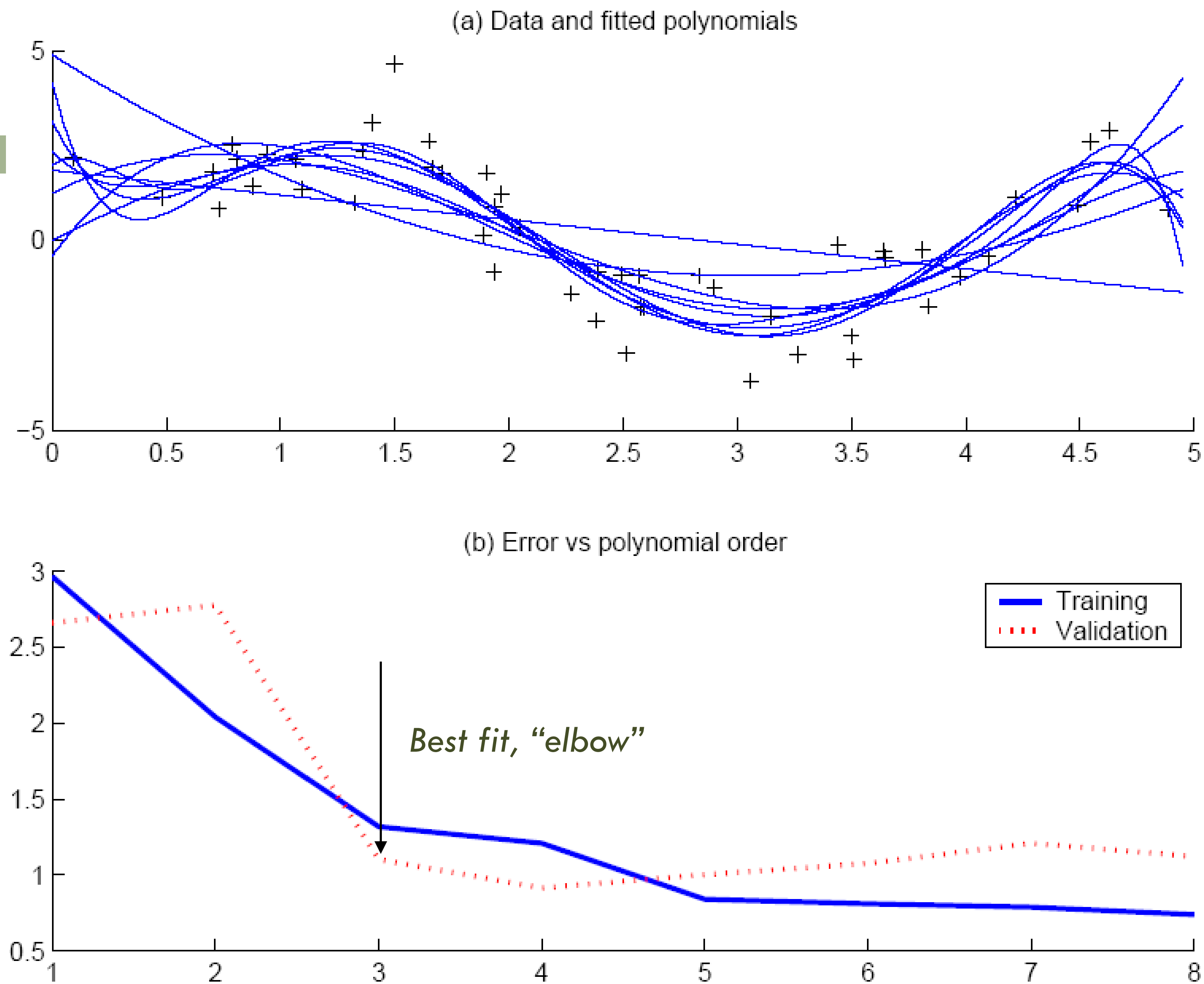
- Bias/Variance dilemma: (Geman et al., 1992)



(a) Function and data
f
(b) Order 1
f
bias
g_i
$\bar{g}$
(c) Order 3
variance
(d) Order 5

Polynomial Regression




(a) Data and fitted polynomials
5
0
−5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
(b) Error vs polynomial order
3
2.5
2
1.5
1
0.5
1
2
3
4
5
6
7
8
Training
Validation
Best fit, "elbow"

Model Selection

- Cross-validation: Measure generalization accuracy by testing on data unused during training
- Regularization: Penalize complex models

  E'=error on data + λ model complexity

  Akaike's information criterion (AIC), Bayesian information criterion (BIC)
- Minimum description length (MDL): Kolmogorov complexity, shortest description of data
- Structural risk minimization (SRM)

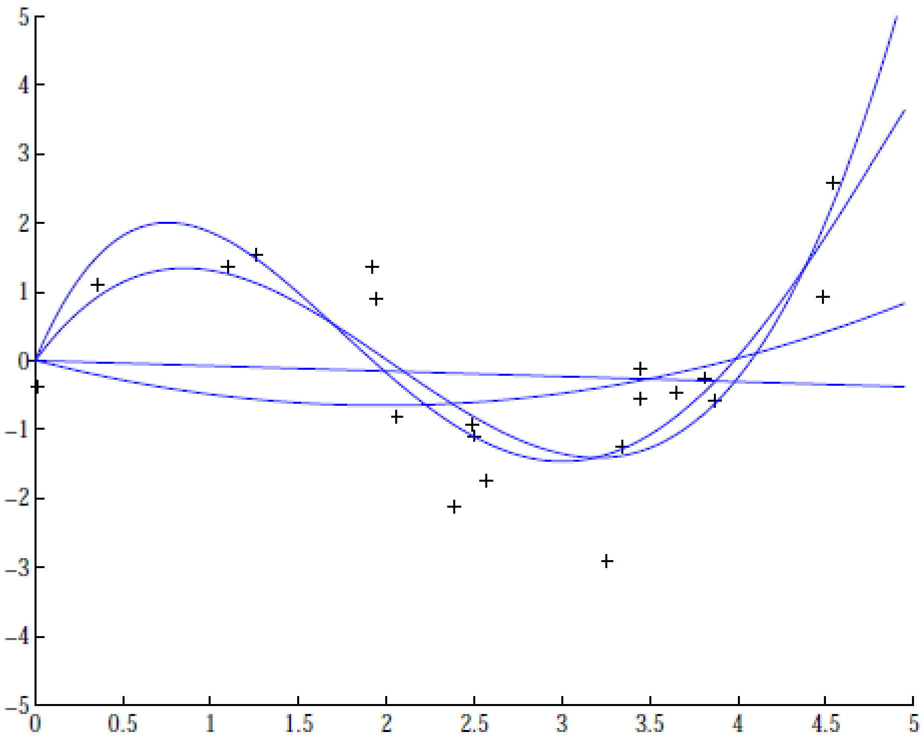
Bayesian Model Selection

- Prior on models, $p(\text{model})$

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

- Regularization, when prior favors simpler models
- Bayes, MAP of the posterior, $p(\text{model} \mid \text{data})$
- Average over a number of models with high posterior (voting, ensembles: Chapter 17)

Regression example



Coefficients increase in magnitude as order increases:

1: [-0.0769, 0.0016]

2: [0.1682, -0.6657, 0.0080]

3: [0.4238, -2.5778, 3.4675, -0.0002

4: [-0.1093, 1.4356, -5.5007, 6.0454, -0.0019]

Regularization (L2): $E(\mathbf{w} \mid \mathcal{X}) = \frac{1}{2} \sum_{t=1}^{N} \left[r^t - g\left(x^t \mid \mathbf{w}\right) \right]^2 + \lambda \sum_i w_i^2$

CHAPTER 5:

MULTİVARİATE METHODS

Multivariate Data

- Multiple measurements (sensors)
- *d* inputs/features/attributes: *d*-variate
- *N* instances/observations/examples

$$\mathbf{X} = \begin{bmatrix} X_1^1 & X_2^1 & \cdots & X_d^1 \\ X_1^2 & X_2^2 & \cdots & X_d^2 \\ \vdots & & & \\ X_1^N & X_2^N & \cdots & X_d^N \end{bmatrix}$$

Multivariate Parameters

$$\text{Mean}: E[\mathbf{x}] = \boldsymbol{\mu} = [\mu_1, \ldots, \mu_d]^T$$

$$\text{Covariance}: \sigma_{ij} \equiv \text{Cov}(X_i, X_j)$$

$$\text{Correlation}: \text{Corr}(X_i, X_j) \equiv \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

$$\Sigma \equiv \text{Cov}(\mathbf{X}) = E\left[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T\right] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

Parameter Estimation

Sample mean $\mathbf{m}: m_i = \frac{\sum_{t=1}^{N} x_i^t}{N}, i = 1, \ldots, d$

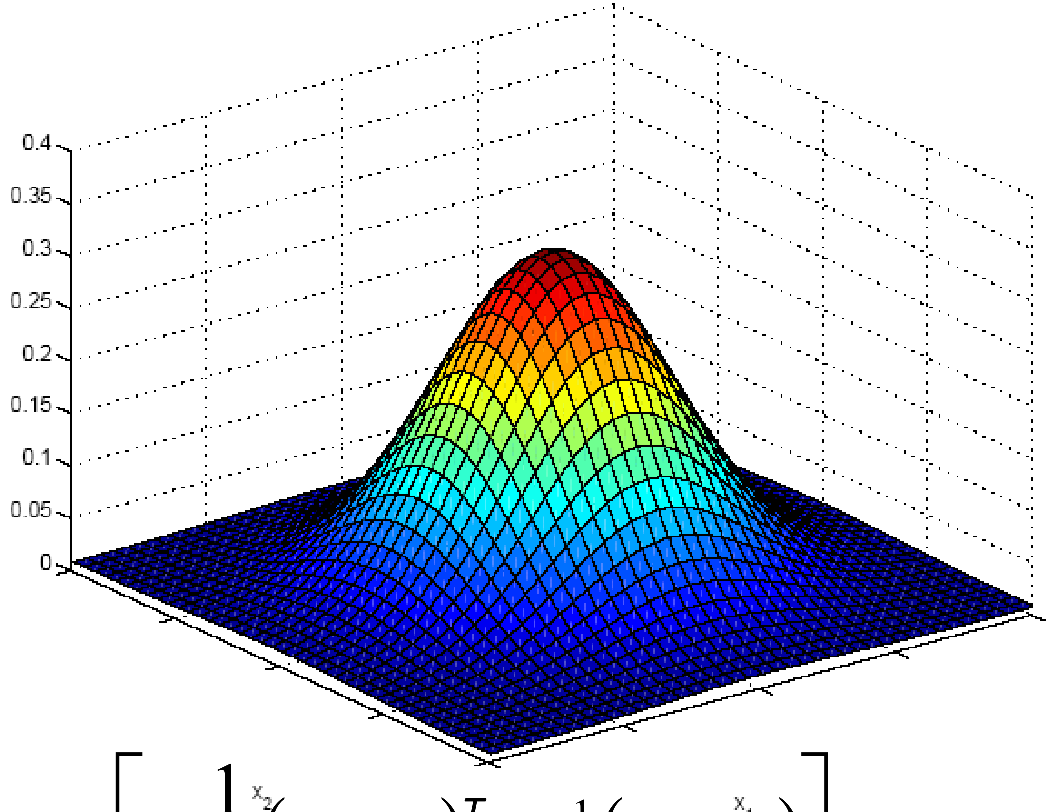
Covariance matrix $\mathbf{S}: s_{ij} = \frac{\sum_{t=1}^{N} \left(x_i^t - m_i\right)\left(x_j^t - m_j\right)}{N}$

Correlation matrix $\mathbf{R}: r_{ij} = \frac{s_{ij}}{s_i s_j}$

Estimation of Missing Values

- What to do if certain instances have missing attributes?
- Ignore those instances: not a good idea if the sample is small
- Use ‘missing’ as an attribute: may give information
- Imputation: Fill in the missing value
  - Mean imputation: Use the most likely value (e.g., mean)
  - Imputation by regression: Predict based on other attributes

Multivariate Normal Distribution



$$\mathbf{x} \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

Multivariate Normal Distribution

- Mahalanobis distance: $(\boldsymbol{x}-\boldsymbol{\mu})^T \sum^{-1} (\boldsymbol{x}-\boldsymbol{\mu})$

  measures the distance from $\boldsymbol{x}$ to $\boldsymbol{\mu}$ in terms of $\sum$ (normalizes for difference in variances and correlations)

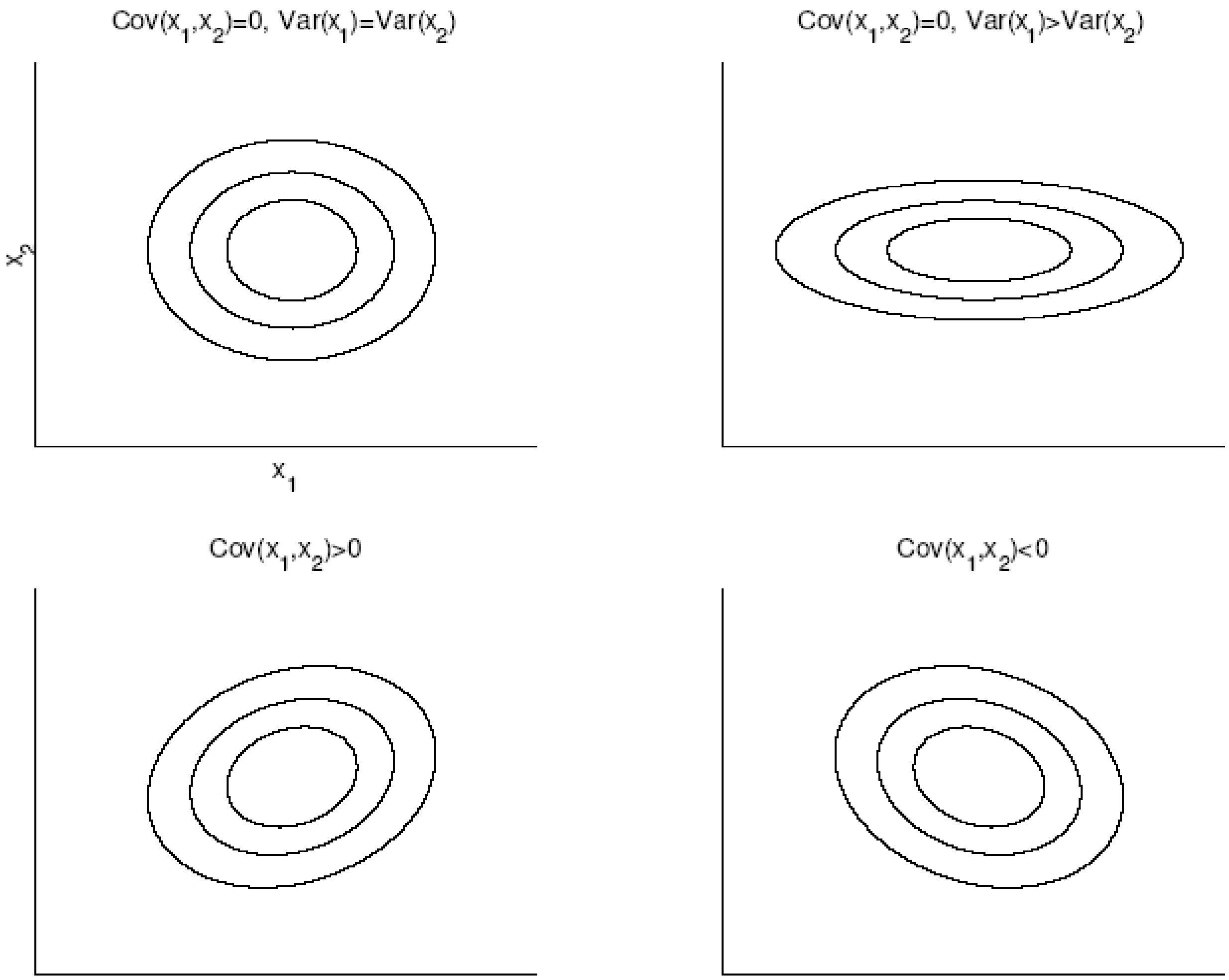
- Bivariate: $d = 2$

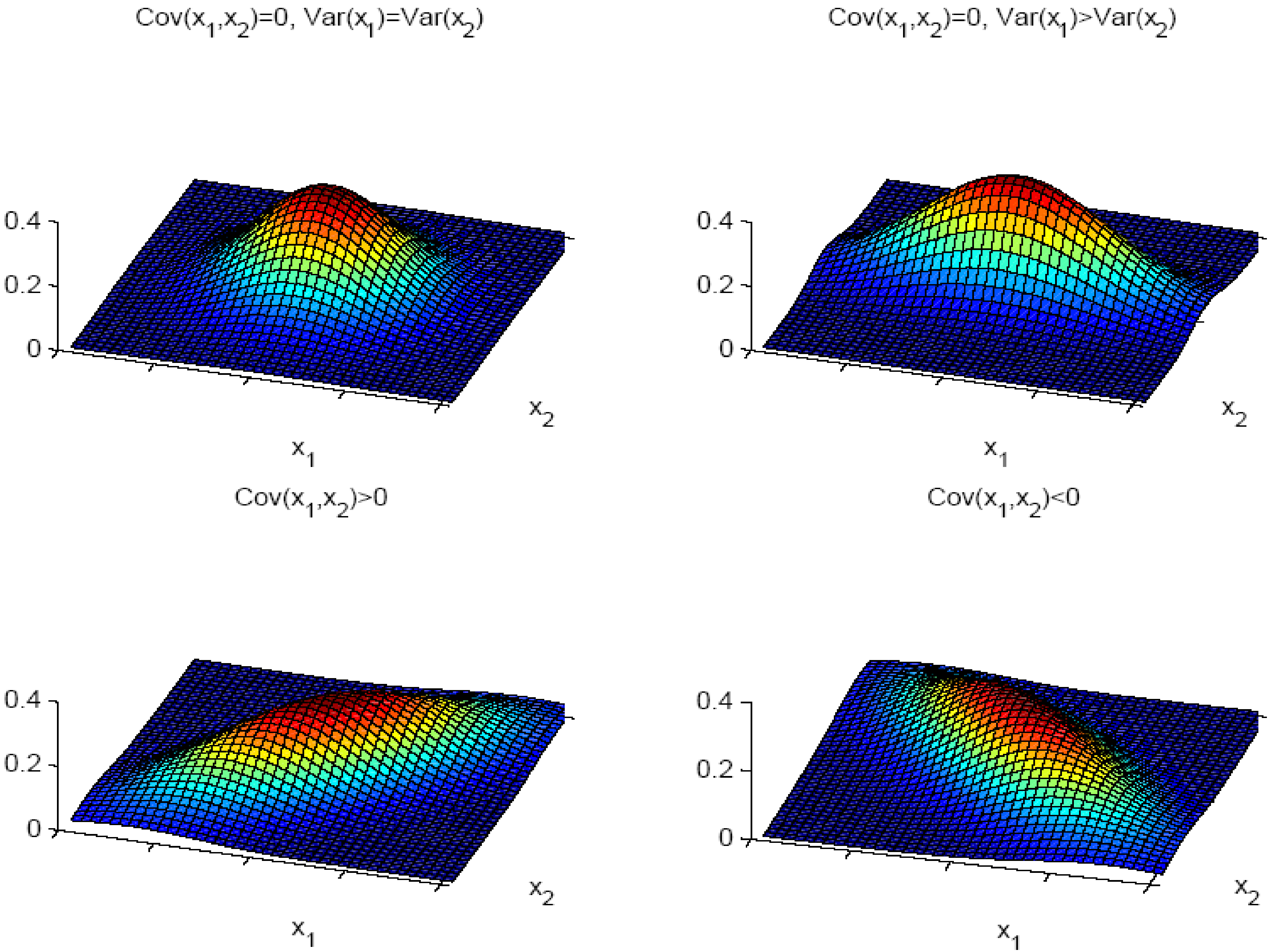
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right)\right]$$

$$z_i = (x_i - \mu_i)/\sigma_i$$

Bivariate Normal





Cov(x_1,x_2)=0, Var(x_1)=Var(x_2)
0.4
0.2
0
x_1
x_2
Cov(x_1,x_2)=0, Var(x_1)>Var(x_2)
0.4
0.2
0
x_1
x_2
Cov(x_1,x_2)>0
0.4
0.2
0
x_1
x_2
Cov(x_1,x_2)<0
0.4
0.2
0
x_1
x_2

Independent Inputs: Naive Bayes

- If x_i are independent, offdiagonals of $\sum$ are 0, Mahalanobis distance reduces to weighted (by $1/\sigma_i$) Euclidean distance:

$$p(\mathbf{x}) = \prod_{i=1}^{d} p_i(x_i) = \frac{1}{(2\pi)^{d/2} \prod_{i=1}^{d} \sigma_i} \exp\left[-\frac{1}{2} \sum_{i=1}^{d} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \right]$$

- If variances are also equal, reduces to Euclidean distance

Parametric Classification

- If $p(\mathbf{x} \mid C_i) \sim N(\boldsymbol{\mu}_i, \Sigma_i)$

$$p(\mathbf{x} \mid C_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)\right]$$

- Discriminant functions

$$\begin{aligned} g_i(\mathbf{x}) &= \log p(\mathbf{x} \mid C_i) + \log P(C_i) \\ &= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i) + \log P(C_i) \end{aligned}$$

Estimation of Parameters

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$$

$$\mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$$

$$\mathbf{S}_i = \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T}{\sum_t r_i^t}$$

$$g_i(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x} - \mathbf{m}_i) + \log \hat{P}(C_i)$$

Different $\mathbf{S}_i$

- Quadratic discriminant

$$g_i(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}\left(\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{x} - 2\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{m}_i + \mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i\right) + \log \hat{P}(C_i)$$

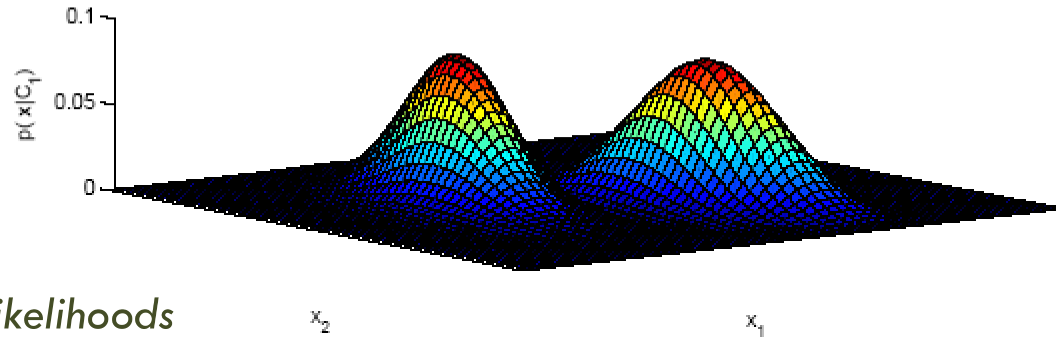
$$= \mathbf{x}^T\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^T\mathbf{x} + w_{i0}$$

where

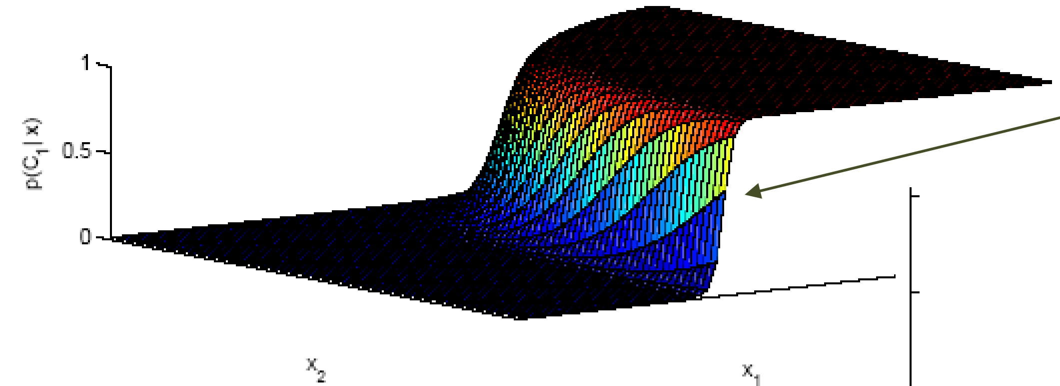
$$\mathbf{W}_i = -\frac{1}{2}\mathbf{S}_i^{-1}$$

$$\mathbf{w}_i = \mathbf{S}_i^{-1}\mathbf{m}_i$$

$$w_{i0} = -\frac{1}{2}\mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i - \frac{1}{2}\log|\mathbf{S}_i| + \log \hat{P}(C_i)$$

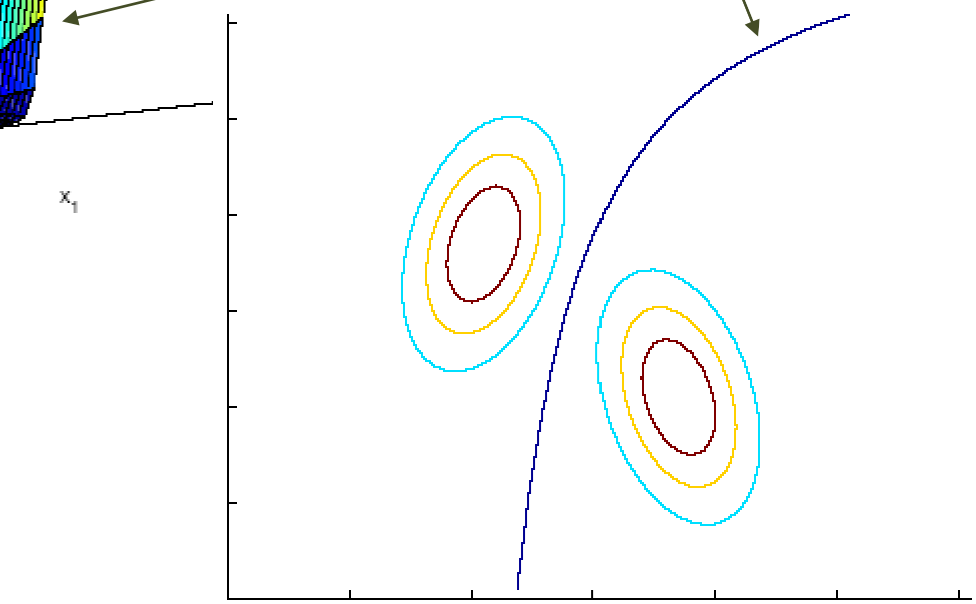


likelihoods



posterior for C_1

discriminant:
P $(C_1 \mid \mathbf{x}) = 0.5$



Common Covariance Matrix S

- Shared common sample covariance **S**

$$\mathbf{S} = \sum_i \hat{P}(C_i)\mathbf{S}_i$$

- Discriminant reduces to

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_i)^T \mathbf{S}^{-1}(\boldsymbol{x} - \boldsymbol{m}_i) + \log \hat{P}(C_i)$$

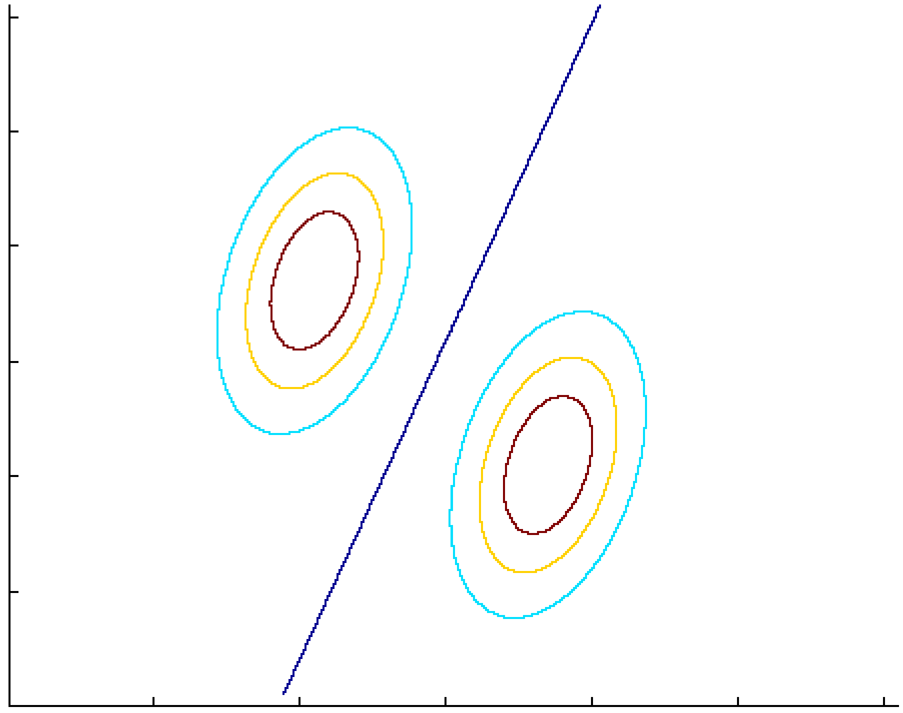
which is a linear discriminant

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

where

$$\mathbf{w}_i = \mathbf{S}^{-1}\mathbf{m}_i \quad w_{i0} = -\frac{1}{2}\mathbf{m}_i^T \mathbf{S}^{-1}\mathbf{m}_i + \log \hat{P}(C_i)$$

Common Covariance Matrix S



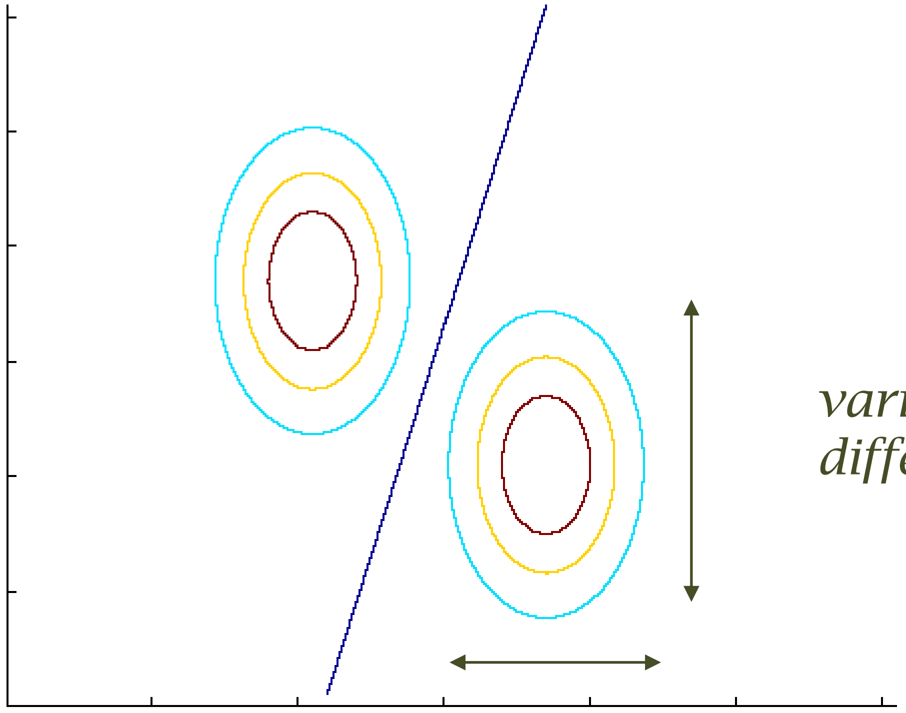
Diagonal S

- When x_j $j = 1,..d$, are independent, $\sum$ is diagonal $p(\mathbf{x}|C_i) = \prod_j p(x_j|C_i)$ (Naive Bayes' assumption)

$$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j^t - m_{ij}}{s_j}\right)^2 + \log \hat{P}(C_i)$$

Classify based on weighted Euclidean distance (in s_j units) to the nearest mean

Diagonal Σ



variances may be different

Diagonal S, equal variances

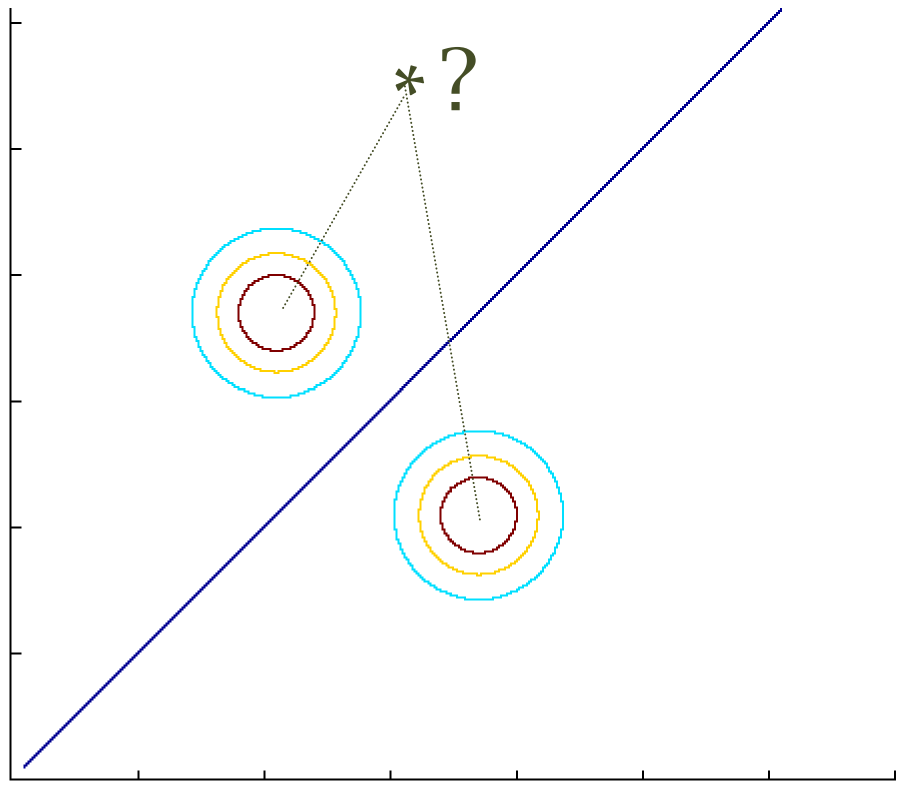
- Nearest mean classifier: Classify based on Euclidean distance to the nearest mean

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\mathbf{m}_i\|^2}{2s^2} + \log \hat{P}(C_i)$$

$$= -\frac{1}{2s^2}\sum_{j=1}^{d}\left(x_j^t - m_{ij}\right)^2 + \log \hat{P}(C_i)$$

- Each mean can be considered a prototype or template and this is template matching

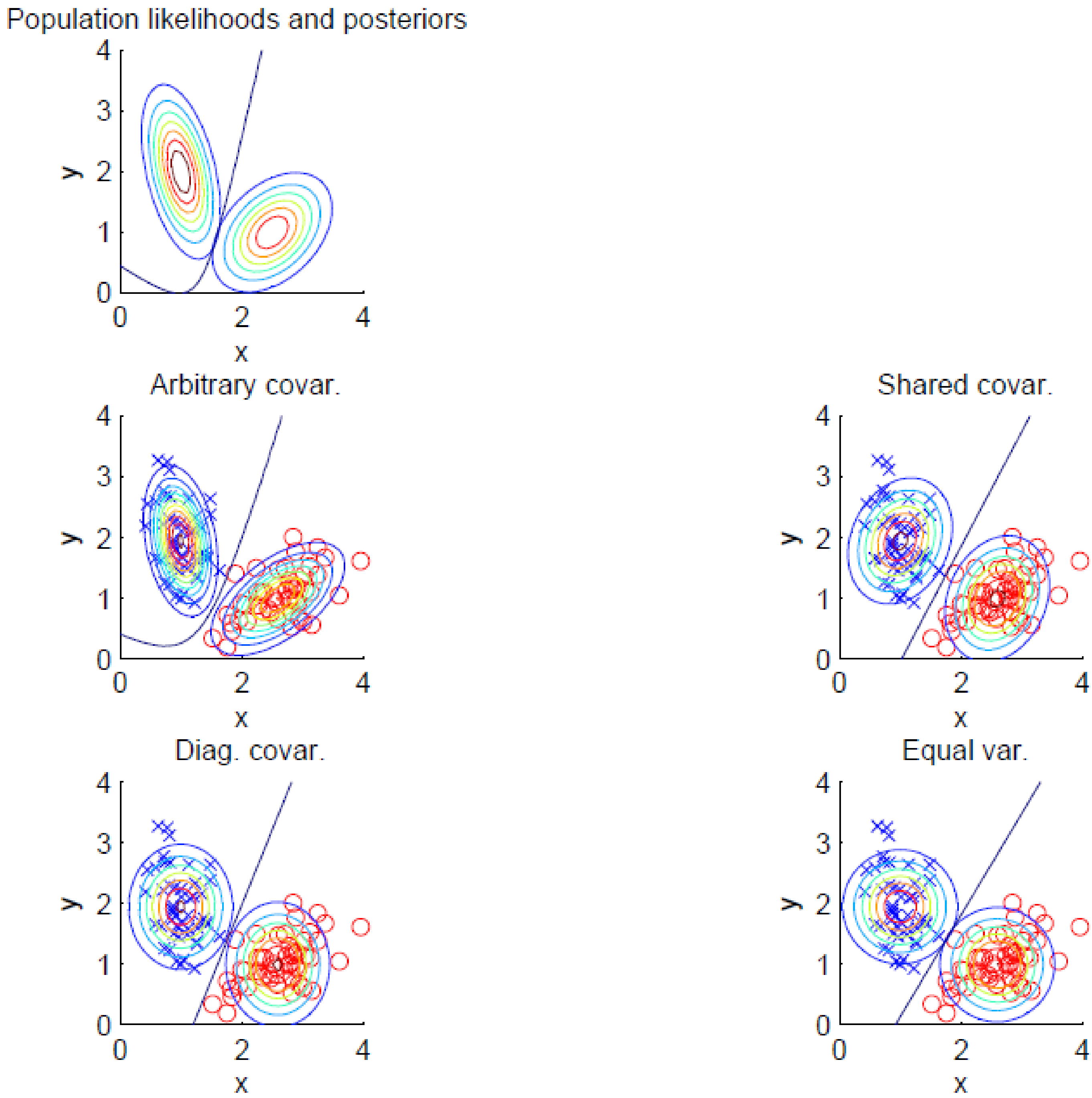
Diagonal **S**, equal variances



Model Selection

Assumption	*Covariance matrix*	*No of parameters*
Shared, Hyperspheric	$\mathbf{S}_i=\mathbf{S}=s^2\mathbf{I}$	1
Shared, Axis-aligned	$\mathbf{S}_i=\mathbf{S}$, with $s_{ij}=0$	d
Shared, Hyperellipsoidal	$\mathbf{S}_i=\mathbf{S}$	$d(d+1)/2$
Different, Hyperellipsoidal	$\mathbf{S}_i$	$K\ d(d+1)/2$

- As we increase complexity (less restricted **S**), bias decreases and variance increases
- Assume simple models (allow some bias) to control variance (regularization)


Population likelihoods and posteriors
Arbitrary covar.
Shared covar.
Diag. covar.
Equal var.
x
y

Discrete Features

- Binary features: $p_{ij} \equiv p(x_j = 1 | C_i)$

  if x_j are independent (Naive Bayes')

$$p(x | C_i) = \prod_{j=1}^{d} p_{ij}^{x_j} (1 - p_{ij})^{(1 - x_j)}$$

  the discriminant is linear

$$g_i(\mathbf{x}) = \log p(\mathbf{x} | C_i) + \log P(C_i)$$
$$= \sum_j \left[x_j \log p_{ij} + (1 - x_j) \log (1 - p_{ij}) \right] + \log P(C_i)$$

  Estimated parameters

$$\hat{p}_{ij} = \frac{\sum_t x_j^t r_i^t}{\sum_t r_i^t}$$

Discrete Features

- Multinomial (1-of-n_j) features: $x_j \in \{v_1, v_2, \ldots, v_{n_j}\}$

$$p_{ijk} \equiv p\left(z_{jk} = 1 \mid C_i\right) = p\left(x_j = v_k \mid C_i\right)$$

  if x_j are independent

$$p(\mathbf{x} \mid C_i) = \prod_{j=1}^{d} \prod_{k=1}^{n_j} p_{ijk}^{z_{jk}}$$

$$g_i(\mathbf{x}) = \sum_j \sum_k z_{jk} \log p_{ijk} + \log P(C_i)$$

$$\hat{p}_{ijk} = \frac{\sum_t z_{jk}^t r_i^t}{\sum_t r_i^t}$$

Multivariate Regression

$$r^t = g\left(x^t \mid w_0, w_1, \ldots, w_d\right) + \varepsilon$$

Multivariate linear model

$$w_0 + w_1 x_1^t + w_2 x_2^t + \cdots + w_d x_d^t$$

$$E\left(w_0, w_1, \ldots, w_d \mid \mathcal{X}\right) = \frac{1}{2}\sum_t \left[r^t - w_0 - w_1 x_1^t - \cdots - w_d x_d^t\right]^2$$

Multivariate polynomial model:

Define new higher-order variables

$z_1 = x_1$, $z_2 = x_2$, $z_3 = x_1^2$, $z_4 = x_2^2$, $z_5 = x_1 x_2$

and use the linear model in this new ***z*** space

(basis functions, kernel trick: Chapter 13)

CHAPTER 13:

KERNEL MACHİNES

Kernel Machines

- Discriminant-based: No need to estimate densities first
- Define the discriminant in terms of support vectors
- The use of kernel functions, application-specific measures of similarity
- No need to represent instances as vectors
- Convex optimization problems with a unique solution

Optimal Separating Hyperplane

$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_t \text{ where } r^t = \begin{cases} +1 & \text{if } \mathbf{x}^t \in C_1 \\ -1 & \text{if } \mathbf{x}^t \in C_2 \end{cases}$$

find $\mathbf{w}$ and w_0 such that

$\mathbf{w}^T \mathbf{x}^t + w_0 \geq +1$ for $r^t = +1$

$\mathbf{w}^T \mathbf{x}^t + w_0 \leq +1$ for $r^t = -1$

which can be rewritten as

$r^t \left(\mathbf{w}^T \mathbf{x}^t + w_0 \right) \geq +1$

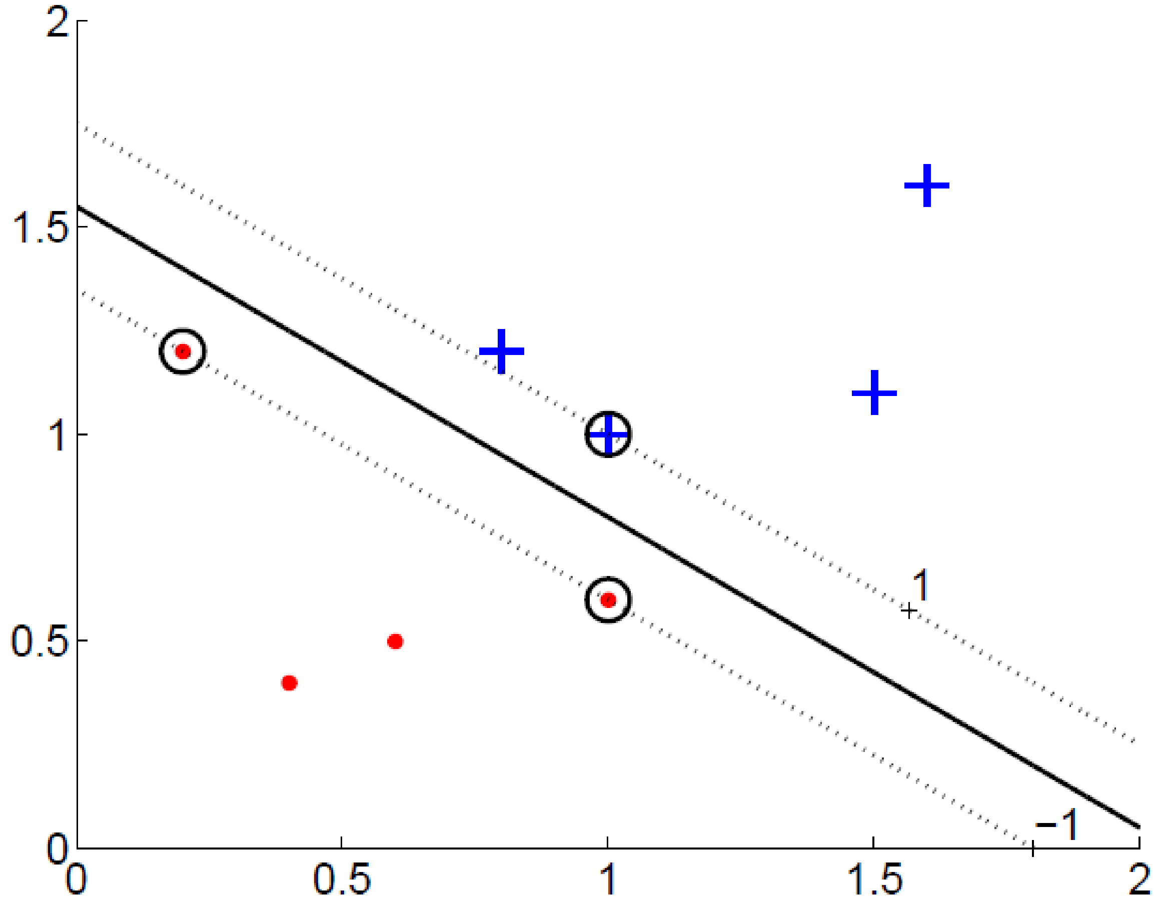
(Cortes and Vapnik, 1995; Vapnik, 1995)

Margin

- Distance from the discriminant to the closest instances on either side
- Distance of x to the hyperplane is $\dfrac{\left|\mathbf{w}^T\mathbf{x}^t + w_0\right|}{\|\mathbf{w}\|}$
- We require $\dfrac{r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right)}{\|\mathbf{w}\|} \geq \rho, \forall t$
- For a unique sol'n, fix $\rho||w||=1$, and to max margin

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1, \forall t$$

Margin



$$\min \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1, \forall t$$

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t \left[r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) - 1\right]$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t r^t \left(\mathbf{w}^T\mathbf{x}^t + w_0\right) + \sum_{t=1}^{N} \alpha^t$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{t=1}^{N} \alpha^t r^t \mathbf{x}^t$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_{t=1}^{N} \alpha^t r^t = 0$$

$$L_d = \frac{1}{2}\left(\mathbf{w}^T\mathbf{w}\right) - \mathbf{w}^T \sum_t \alpha^t r^t \mathbf{x}^t - w_0 \sum_t \alpha^t r^t + \sum_t \alpha^t$$

$$= -\frac{1}{2}\left(\mathbf{w}^T\mathbf{w}\right) + \sum_t \alpha^t$$

$$= -\frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s \left(\mathbf{x}^t\right)^T \mathbf{x}^s + \sum_t \alpha^t$$

subject to $\sum_{t} \alpha^t r^t = 0$ and $\alpha^t \geq 0, \forall t$

Most α^t are 0 and only a small number have $\alpha^t > 0$; they are the support vectors

Soft Margin Hyperplane

- Not linearly separable

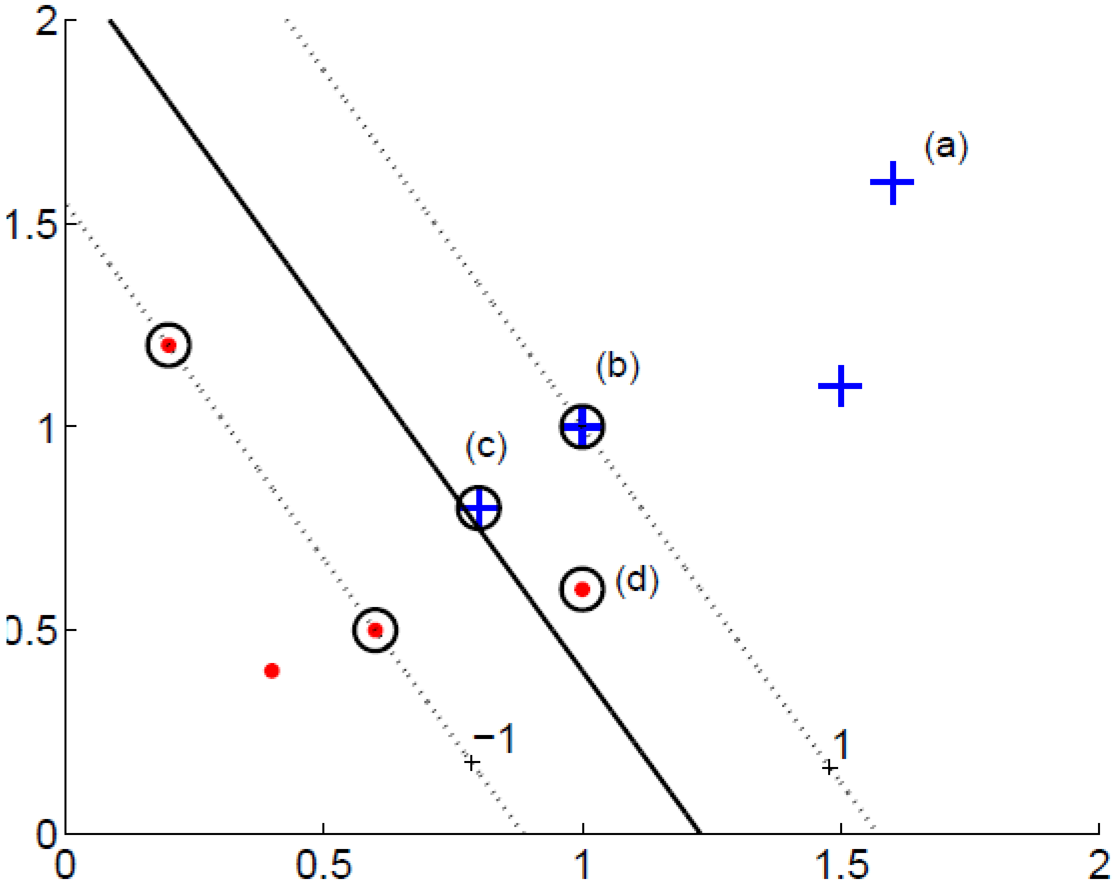
$$r^t\left(\mathbf{w}^T x^t + w_0\right) \geq 1 - \xi^t$$

- Soft error

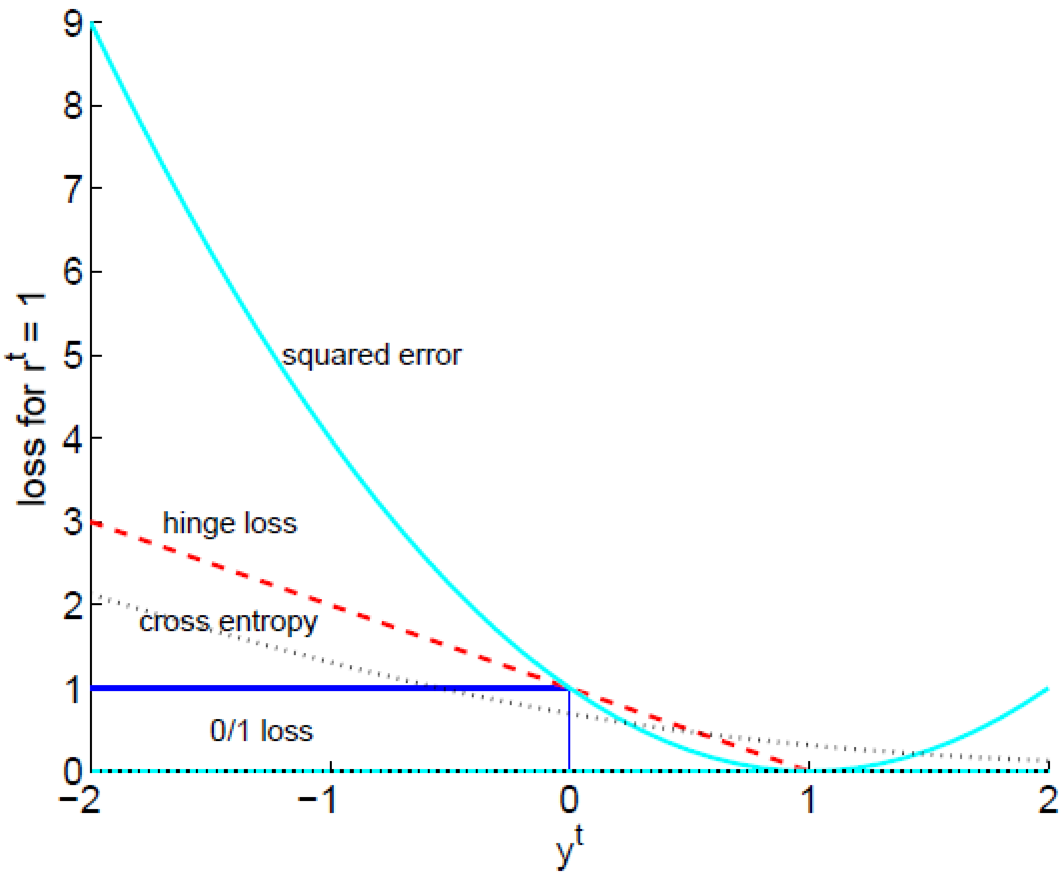
$$\sum_t \xi^t$$

- New primal is

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t - \sum_t \alpha^t \left[r^t\left(\mathbf{w}^T x^t + w_0\right) - 1 + \xi^t\right] - \sum_t \mu^t \xi^t$$


(a)
(b)
(c)
(d)
−1
1
0
0.5
1
1.5
2

Hinge Loss



$$
: \begin{cases} 0 & \text{if } y^t r^t \geq 1 \\ 1 - y^t r^t & \text{otherwise} \end{cases}
$$

ν-SVM

$$\min \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{N}\sum_t \xi^t$$

subject to

$$r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq \rho - \xi^t, \xi^t \geq 0, \rho \geq 0$$

$$L_d = -\frac{1}{2}\sum_{t=1}^{N}\sum_s \alpha^t\alpha^s r^t r^s \left(x^t\right)^T x^s$$

subject to

$$\sum_t \alpha^t r^t = 0, 0 \leq \alpha^t \leq \frac{1}{N}, \sum_t \alpha^t \leq \nu$$

ν controls the fraction of support vectors

Kernel Trick

- Preprocess input **x** by basis functions

$$\boldsymbol{z} = \boldsymbol{\varphi}(\boldsymbol{x}) \qquad g(\boldsymbol{z}) = \boldsymbol{w}^T\boldsymbol{z}$$

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x})$$

- The SVM solution

$$\mathbf{w} = \sum_t \alpha^t r^t \mathbf{z}^t = \sum_t \alpha^t r^t \boldsymbol{\varphi}(\mathbf{x}^t)$$

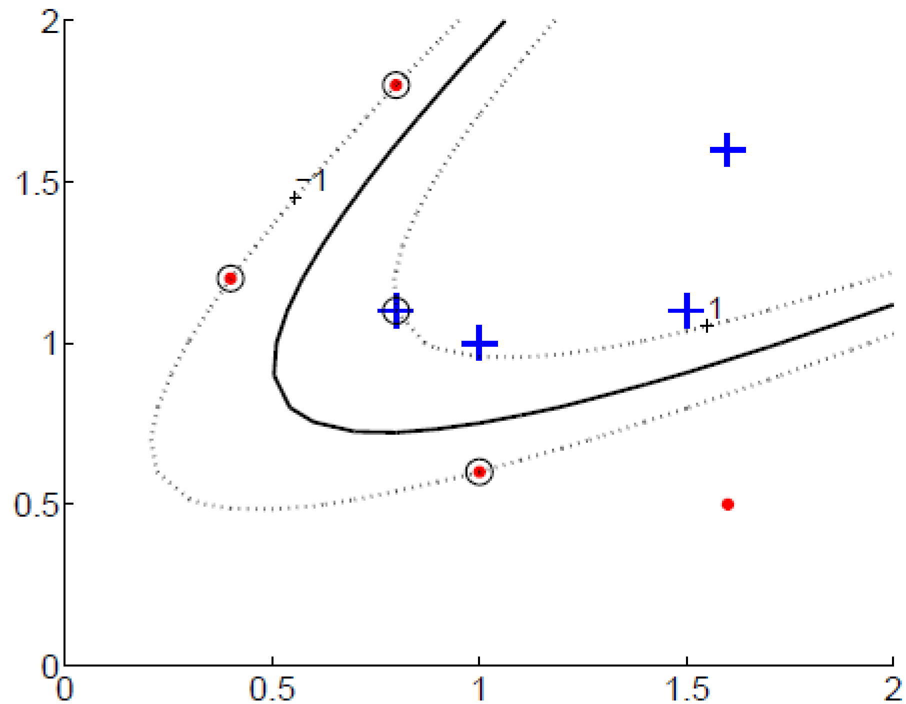
$$g(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) = \sum_t \alpha^t r^t \boxed{\boldsymbol{\varphi}(\mathbf{x}^t)^T \boldsymbol{\varphi}(\mathbf{x})}$$

$$g(\mathbf{x}) = \sum_t \alpha^t r^t \boxed{K(\mathbf{x}^t, \mathbf{x})}$$

Vectorial Kernels

- Polynomials of degree q:

$$K(\mathbf{x}^t, \mathbf{x}) = (\mathbf{x}^T \mathbf{x}^t + 1)^q$$



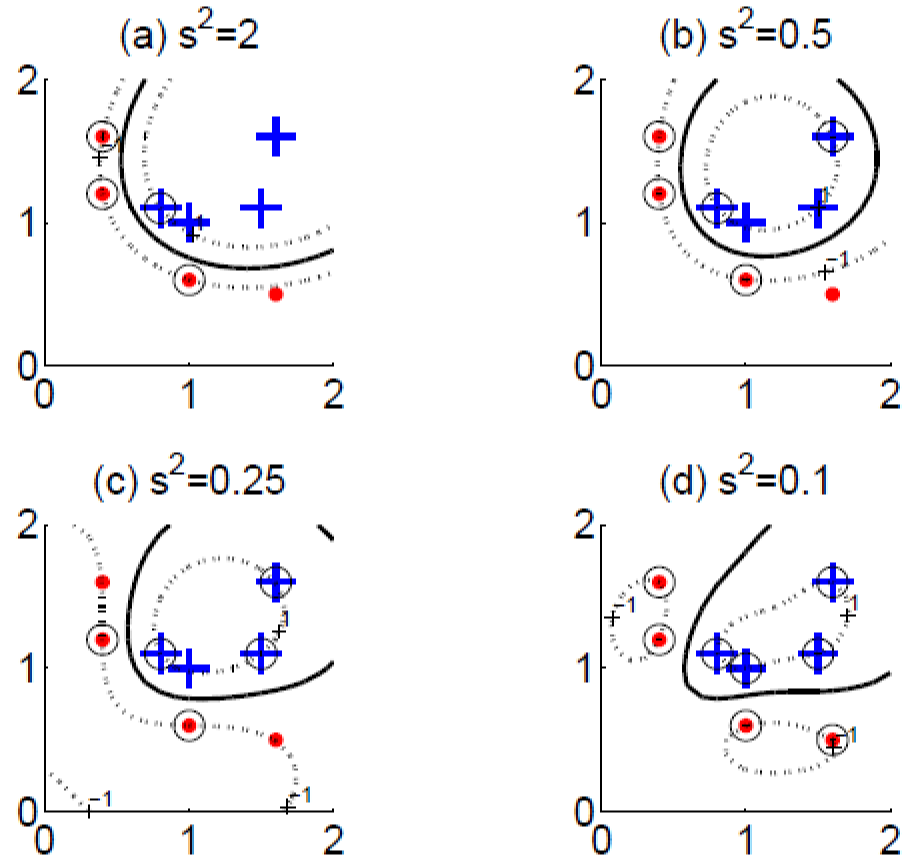
$$\begin{aligned}K(\mathbf{x}, \mathbf{y}) &= (\mathbf{x}^T \mathbf{y} + 1)^2 \\ &= (x_1 y_1 + x_2 y_2 + 1)^2 \\ &= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2\end{aligned}$$

$$\phi(\mathbf{x}) = \left[1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2\right]^T$$

Vectorial Kernels

- Radial-basis functions:

$$K\left(\mathbf{x}^t, \mathbf{x}\right) = \exp\left[-\frac{\left\|\mathbf{x}^t - \mathbf{x}\right\|^2}{2s^2}\right]$$



Defining kernels

- Kernel "engineering"
- Defining good measures of similarity
- String kernels, graph kernels, image kernels, ...
- Empirical kernel map: Define a set of templates $\mathbf{m}_i$ and score function $s(\mathbf{x},\mathbf{m}_i)$

  $\phi(\mathbf{x}^t)=[s(\mathbf{x}^t,\mathbf{m}_1), s(\mathbf{x}^t,\mathbf{m}_2),..., s(\mathbf{x}^t,\mathbf{m}_M)]$

  and

  $K(\mathbf{x},\mathbf{x}^t)=\phi(\mathbf{x})^T \phi(\mathbf{x}^t)$

Multiple Kernel Learning

- Fixed kernel combination

$$K(\mathbf{x},\mathbf{y}) = \begin{cases} cK(\mathbf{x},\mathbf{y}) \\ K_1(\mathbf{x},\mathbf{y}) + K_2(\mathbf{x},\mathbf{y}) \\ K_1(\mathbf{x},\mathbf{y})K_2(\mathbf{x},\mathbf{y}) \end{cases}$$

- Adaptive kernel combination

$$K(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{m} \eta_i K_i(\mathbf{x},\mathbf{y})$$

$$L_d = \sum_t \alpha^t - \frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s \sum_i \eta_i K_i(\mathbf{x}^t, \mathbf{x}^s)$$

$$g(\mathbf{x}) = \sum_t \alpha^t r^t \sum_i \eta_i K_i(\mathbf{x}^t, \mathbf{x})$$

- Localized kernel combination $g(\mathbf{x}) = \sum_t \alpha^t r^t \sum_i \eta_i(\mathbf{x} \mid \theta) K_i(\mathbf{x}^t, \mathbf{x})$

Multiclass Kernel Machines

- 1-vs-all
- Pairwise separation
- Error-Correcting Output Codes (section 17.5)
- Single multiclass optimization

$$\min \frac{1}{2}\sum_{i=1}^{K}\|\mathbf{w}_i\|^2 + C\sum_i\sum_t \xi_i^t$$

subject to

$$\mathbf{w}_{z^t}{}^T\mathbf{x}^t + w_{z^t 0} \geq \mathbf{w}_i{}^T\mathbf{x}^t + w_{i0} + 2 - \xi_i^t, \forall i \neq z^t, \xi_i^t \geq 0$$

SVM for Regression

- Use a linear model (possibly kernelized)

$$f(\mathbf{x})=\mathbf{w}^T\mathbf{x}+w_0$$

- Use the ϵ-sensitive error function

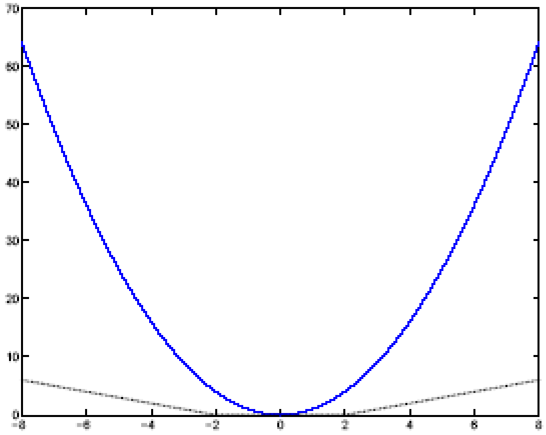
$$e_\varepsilon\left(r^t, f\left(\mathbf{x}^t\right)\right)=\begin{cases}0 & \text{if } \left|r^t-f\left(\mathbf{x}^t\right)\right|<\varepsilon \\ \left|r^t-f\left(\mathbf{x}^t\right)\right|-\varepsilon & \text{otherwise}\end{cases}$$

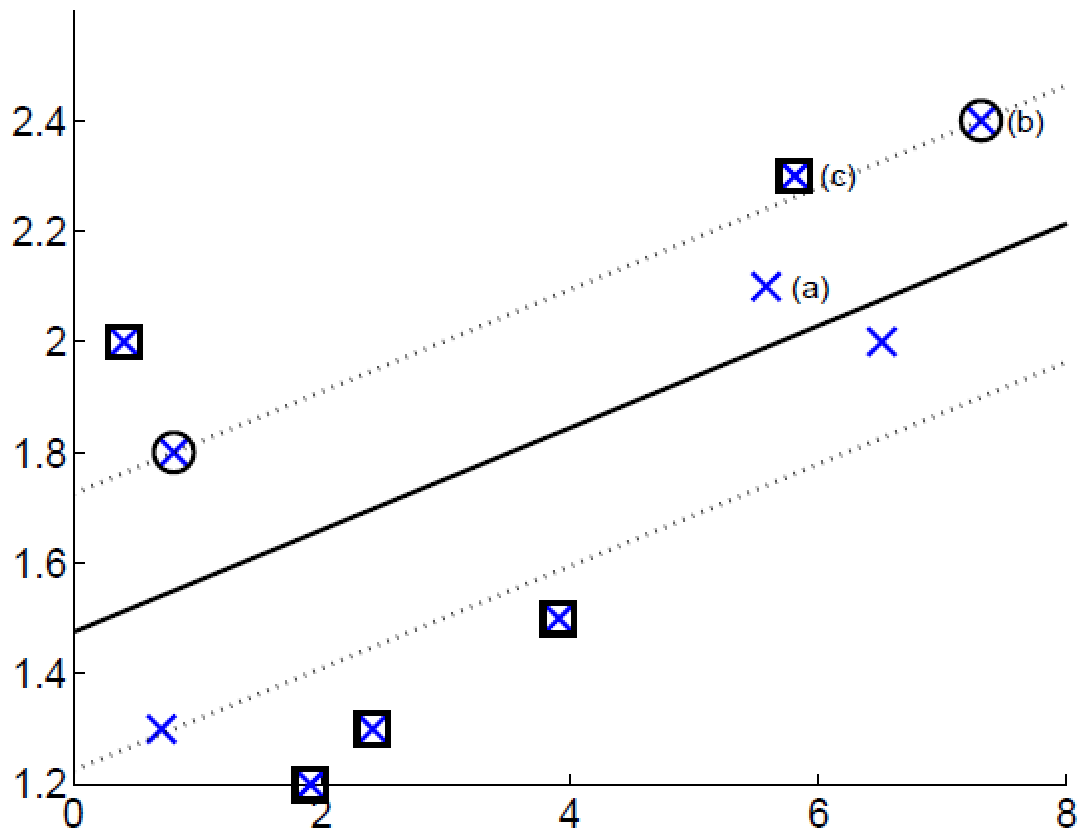
$$\min \frac{1}{2}\|\mathbf{w}\|^2+C\sum_t\left(\xi_+^t+\xi_-^t\right)$$

$$r^t-\left(\mathbf{w}^T\mathbf{x}+w_0\right)\leq\varepsilon+\xi_+^t$$

$$\left(\mathbf{w}^T\mathbf{x}+w_0\right)-r^t\leq\varepsilon+\xi_-^t$$

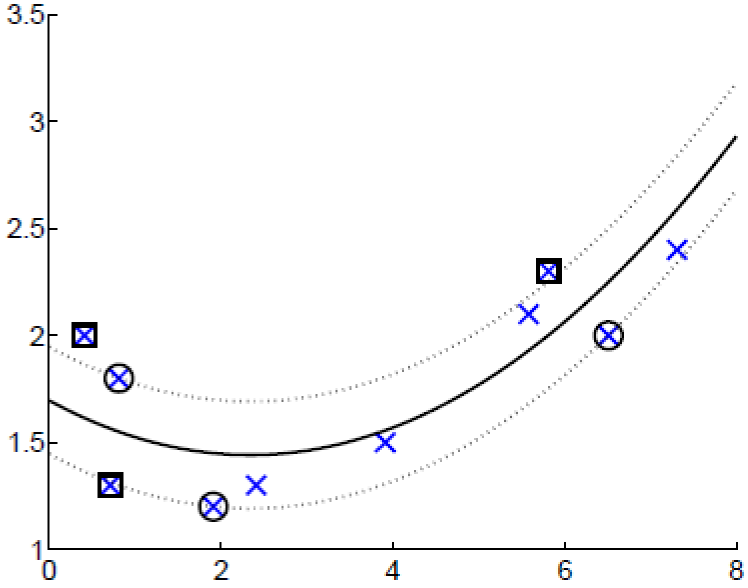
$$\xi_+^t, \xi_-^t\geq 0$$



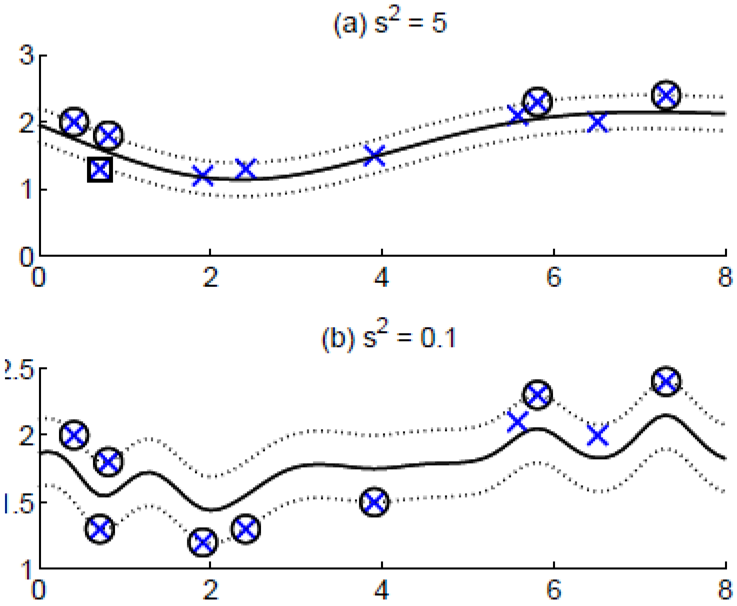

2.4
2.2
2
1.8
1.6
1.4
1.2
0
2
4
6
8
(a)
(b)
(c)

Kernel Regression

- Polynomial kernel



- Gaussian kernel



Kernel Machines for Ranking

- We require not only that scores be correct order but at least +1 unit margin.
- Linear case:

$$\min \frac{1}{2}\|\mathbf{w}_i\|^2 + C\sum_t \xi_i^t$$

subject to

$$\mathbf{w}^T\mathbf{x}^u \geq \mathbf{w}^T\mathbf{x}^v + 1 - \xi^t, \forall t : r^u \prec r^v, \xi_i^t \geq 0$$

One-Class Kernel Machines

- Consider a sphere with center ***a*** and radius *R*

$$\min R^2 + C\sum_t \xi^t$$

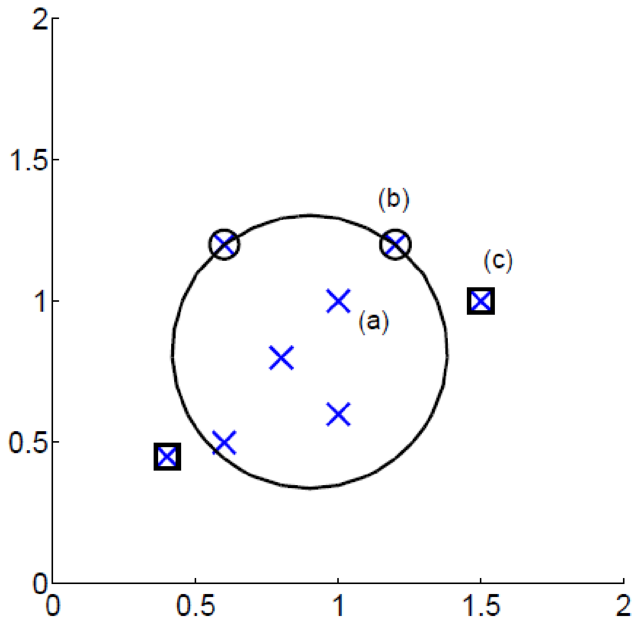
subject to

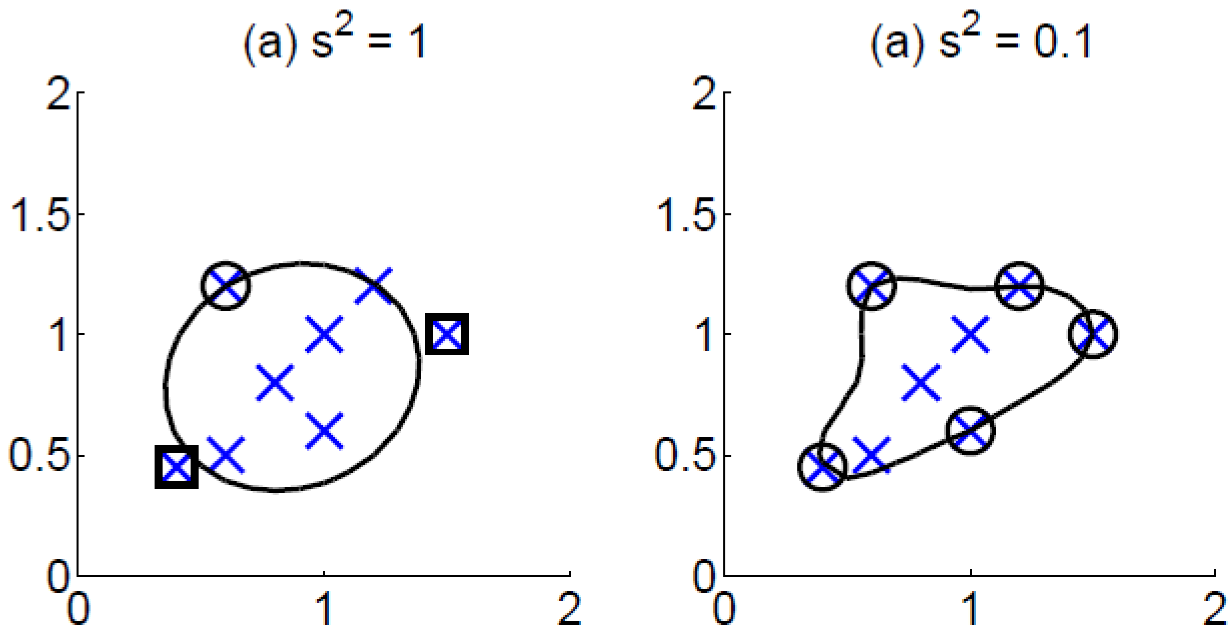
$$\left\|\mathbf{x}^t - a\right\| \leq R^2 + \xi^t, \xi^t \geq 0$$

$$L_d = \sum_t \alpha^t \boxed{\left(x^t\right)^T x^s} - \sum_{t=1}^{N}\sum_s \alpha^t \alpha^s r^t r^s \boxed{\left(x^t\right)^T x^s}$$

subject to

$$0 \leq \alpha^t \leq C, \sum_t \alpha^t = 1$$




(a) s2 = 1
(a) s2 = 0.1
2
1.5
1
0.5
0
0
1
2
2
1.5
1
0.5
0
0
1
2

Large Margin Nearest Neighbor

- Learns the matrix **M** of Mahalanobis metric

  $D(\mathbf{x}^i, \mathbf{x}^j)=(\mathbf{x}^i-\mathbf{x}^j)^T\mathbf{M}(\mathbf{x}^i-\mathbf{x}^j)$

- For three instances *i*, *j*, and *l*, where *i* and *j* are of the same class and *l* different, we require

  $D(\mathbf{x}^i, \mathbf{x}^l) > D(\mathbf{x}^i, \mathbf{x}^j)+1$

  and if this is not satisfied, we have a slack for the difference and we learn M to minimize the sum of such slacks over all *i*,*j*,*l* triples (*j* and *l* being one of *k* neighbors of *i*, over all *i*)

Learning a Distance Measure

- LMNN algorithm (Weinberger and Saul 2009)

$$(1-\mu)\sum_{i,j}\mathcal{D}(\mathbf{x}^i,\mathbf{x}^j)+\mu\sum_{i,j,l}(1-y_{il})\xi_{ijl}$$

subject to

$$\begin{aligned}\mathcal{D}(\mathbf{x}^i,\mathbf{x}^l) &\geq \mathcal{D}(\mathbf{x}^i,\mathbf{x}^j)+1-\xi^{ijl}, \text{ if } \boldsymbol{r}^i=\boldsymbol{r}^j \text{ and } \boldsymbol{r}^i\neq\boldsymbol{r}^l\\ \xi^{ijl} &\geq 0\end{aligned}$$

- LMCA algorithm (Torresani and Lee 2007) uses a similar approach where $\mathbf{M}=\mathbf{L}^T\mathbf{L}$ and learns $\mathbf{L}$

Kernel Dimensionality Reduction

- Kernel PCA does PCA on the kernel matrix (equal to canonical PCA with a linear kernel)
- Kernel LDA, CCA