Introduction to ML

October 4^{th}, 2021

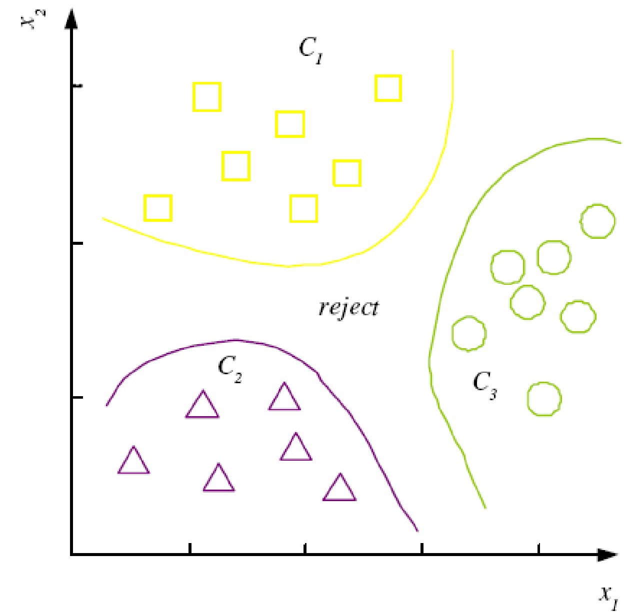
Discriminant Functions

Classification rule: pick one such function that maximizes

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

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

Three different discriminant functions



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

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

Discriminant functions:
Divides feature space into K region
For those inputs x, find the function that give the largest value (use that function to carve out a region

Association Rules

- Association rule: $X \rightarrow Y$
- *People who buy/click/visit/enjoy* ***X*** *are also likely to buy/click/visit/enjoy* ***Y****.*
  - Dependency, it's not cause and effect but good enough for making a business decision

- A rule implies association, not necessarily causation.

Measures for assessing association

Association measures

Joint probability,
Make this large (increase the basis)
Significance of the rule
Useless is this number is low

- Support ($X \rightarrow Y$):

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

- Confidence ($X \rightarrow Y$):

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

- conditional probability
- Should be larger than P(Y)
- **Strength** of the association rule

- Lift ($X \rightarrow Y$):

$$= \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

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

- If independent, this is 1
- If lift>1 X makes Y more likely

Example

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

SOLUTION:

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

In making a decision, take 'high support + high confidence rule"

CHAPTER 4:

Parametric Methods

Statistics based ML method

Why do we need this?

Using Bayes Decision Theory for Classification

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

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

The distribution function of P(x) when X is coming from a class C_i

Need this probability to be used as discriminant function

Collect data
Use training set
Learn the discriminant function (essentially estimate the parameters)

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

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

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

  How about Bernoulli? Binomial?

All that is required to define a probability distribution

Maximum Likelihood Estimator

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

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

Estimate parameter of the model using X with this criterion

Joint factors to product (assume iid samples)

- 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}

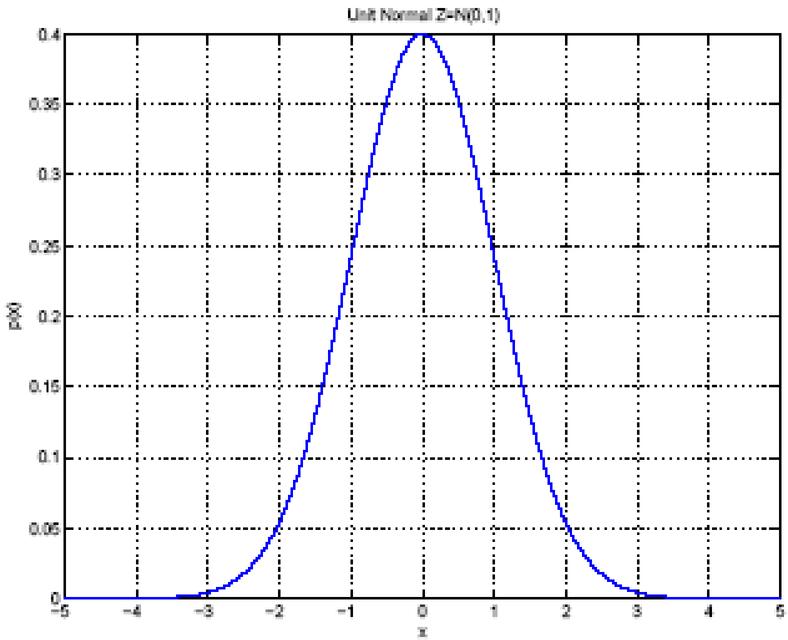
  $P(x_1, x_2, \ldots, x_K) = \prod_i p_i^{x_i}$

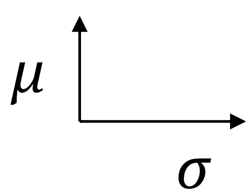
  $$\mathcal{L}(p_1, p_2, \ldots, p_K | \mathcal{X}) = \log \prod_t \prod_i p_i^{x_i^t}$$

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

Like average

Gaussian (Normal) Distribution -Used for continuous-value data





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

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

- MLE for μ and σ^2:

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

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

Bias and Variance

We derive a function to estimate the parameter values of distribution

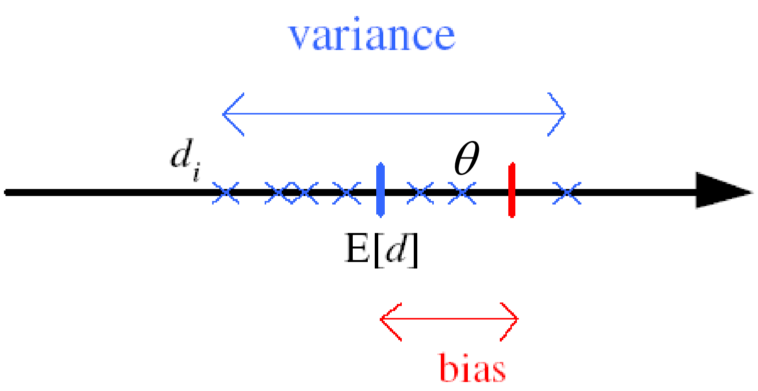
They are a function of rvs, also a rv by itself, understand the property of such a learner is important

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:

$r\,(d,\theta) = E\,[(d-\theta)^2]$

$= (E\,[d] - \theta)^2 + E\,[(d - E\,[d])^2]$

$=$ Bias^2 + Variance

Errors in our estimator consists of two things:

1. Bias^2
2. Variance

Take home thinking: what would minimize this?

Bayes' Estimator
-different from MLE

Parameter of the function is a random variable with a distribution

- Treat $\boldsymbol{\vartheta}$ as a random var with prior $p\ (\boldsymbol{\vartheta})$
- 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) estmation:

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

Find the **conditional expected** value of the parameter random variable given collected data

Bayes' Estimator: Example

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

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

Weighted average between mean of the prior and the mean of the data distribution

What happens when N is very large?
(you learn everything from data)

Parametric Classification

-once done (learning), you also obtain the actual 'discriminant' function that can be used for classification

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

or

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

This is your discriminant function for classification

$$\boxed{\begin{aligned} p(x \mid C_i) &= \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^{\ 2}}\right] \\ g_i(x) &= -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^{\ 2}} + \log P(C_i) \end{aligned}}$$

The actual implementation when using Gaussian distribution

An example of learning for two class problem

- 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

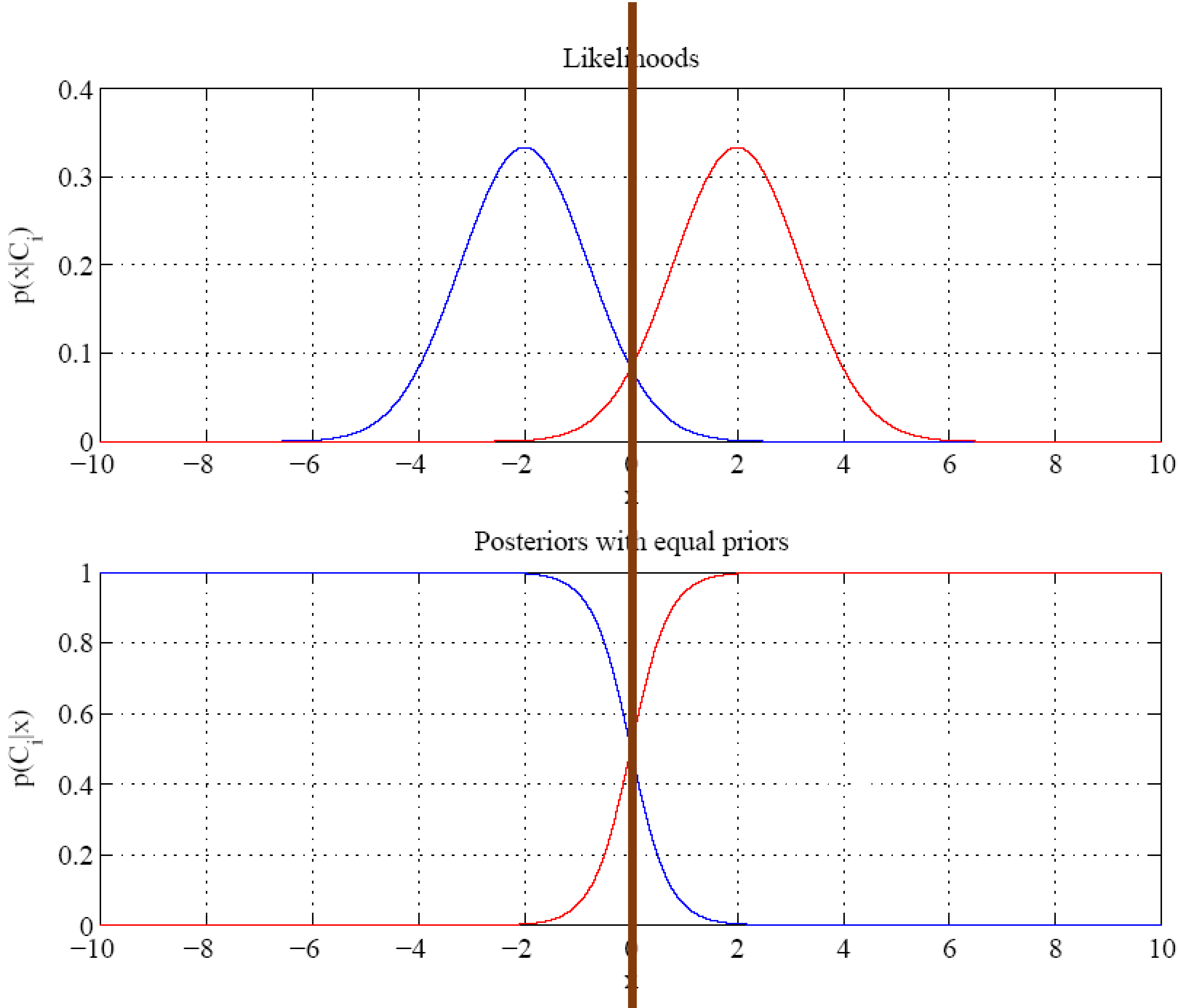
$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \quad s_i^2 = \frac{\sum_t \left(x^t - m_i\right)^2 r_i^t}{\sum_t r_i^t}$$

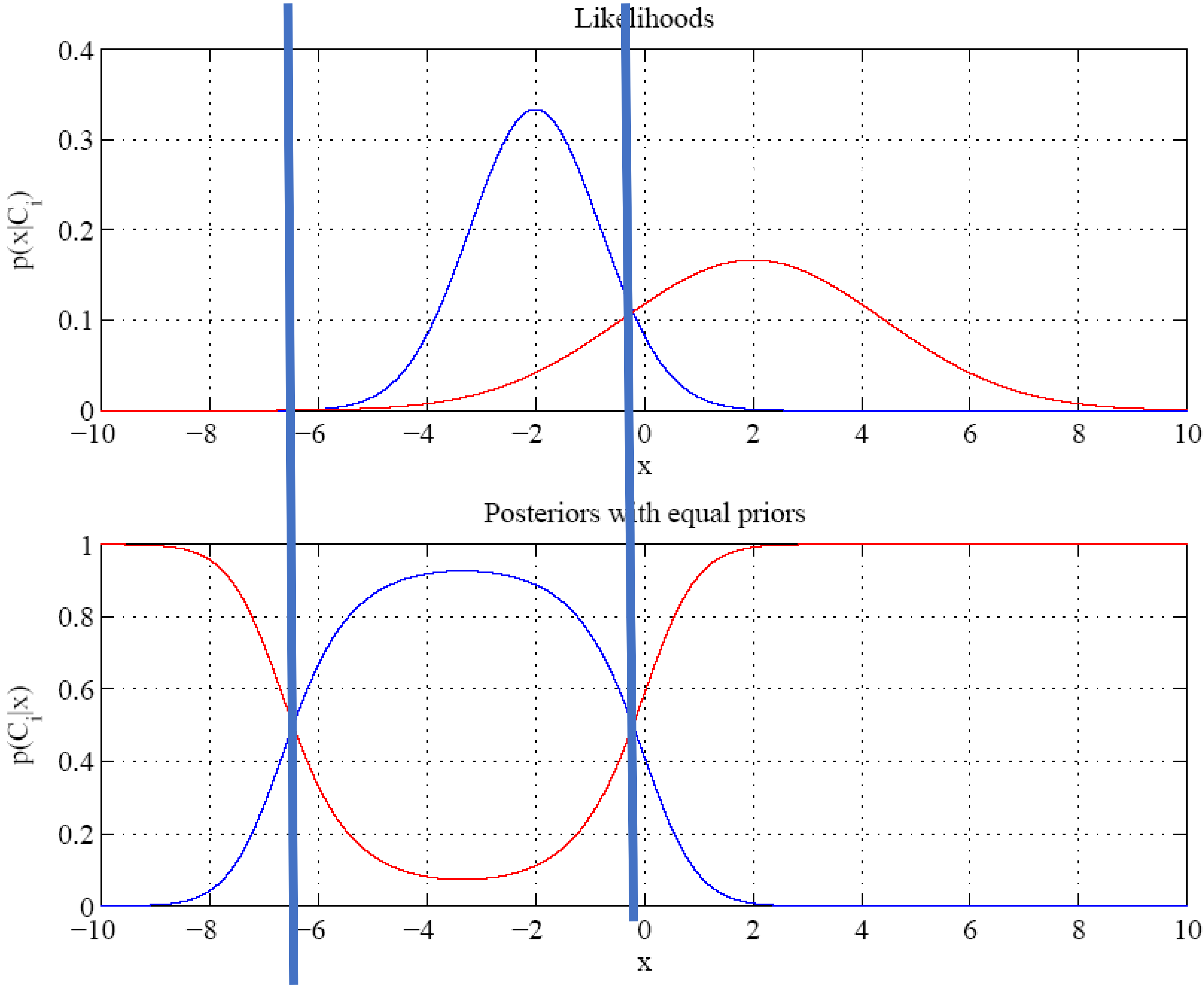
- Discriminant

Once done, use the following function on test set

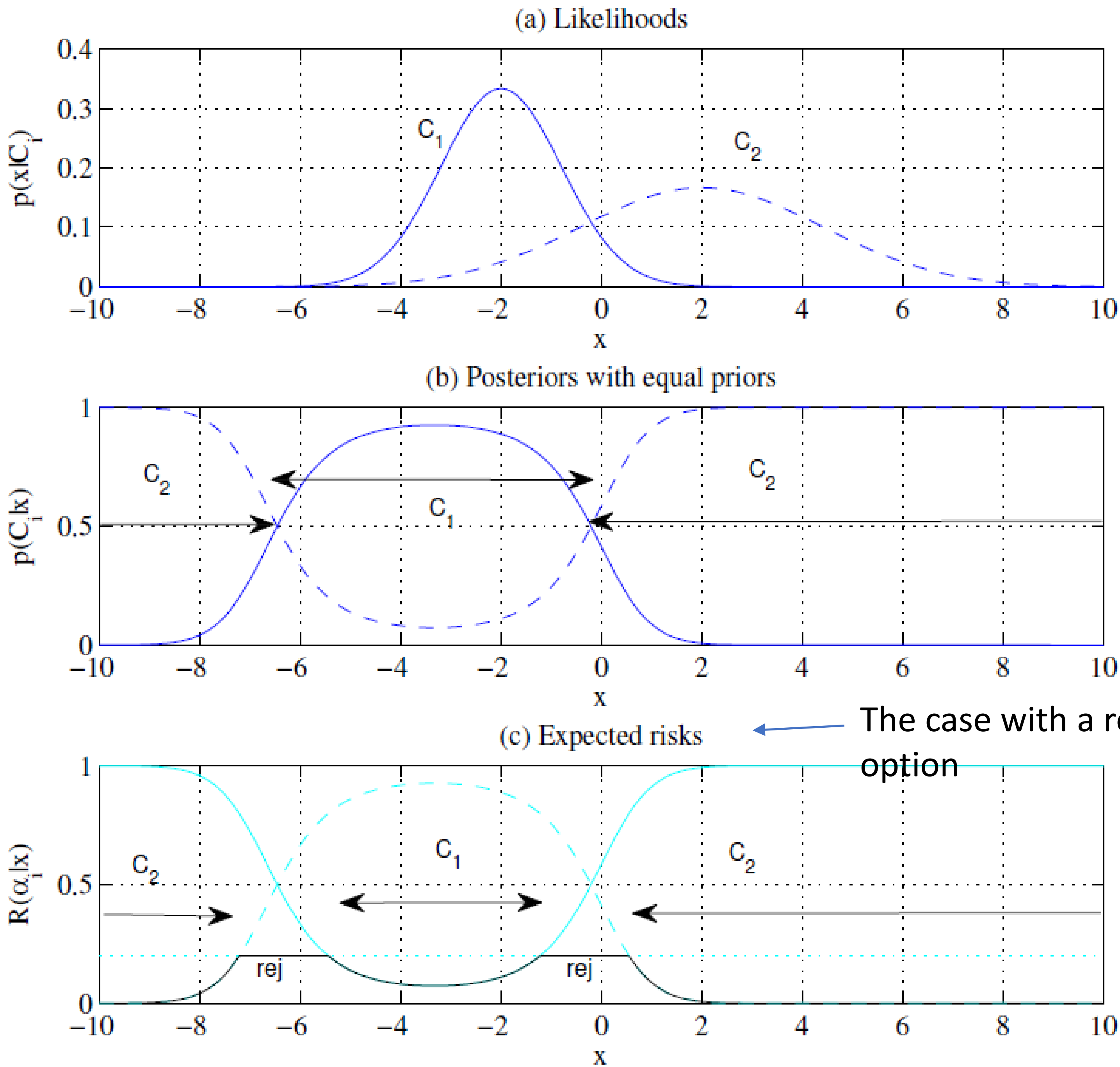
$$g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{\left(x - m_i\right)^2}{2s_i^2} + \log \hat{P}(C_i)$$

Input x (test sample) for each i, look at the class label i, for the one that is max, pick that class i


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



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



The case with a reject option

Regression

measurement

Input x, f is what we want

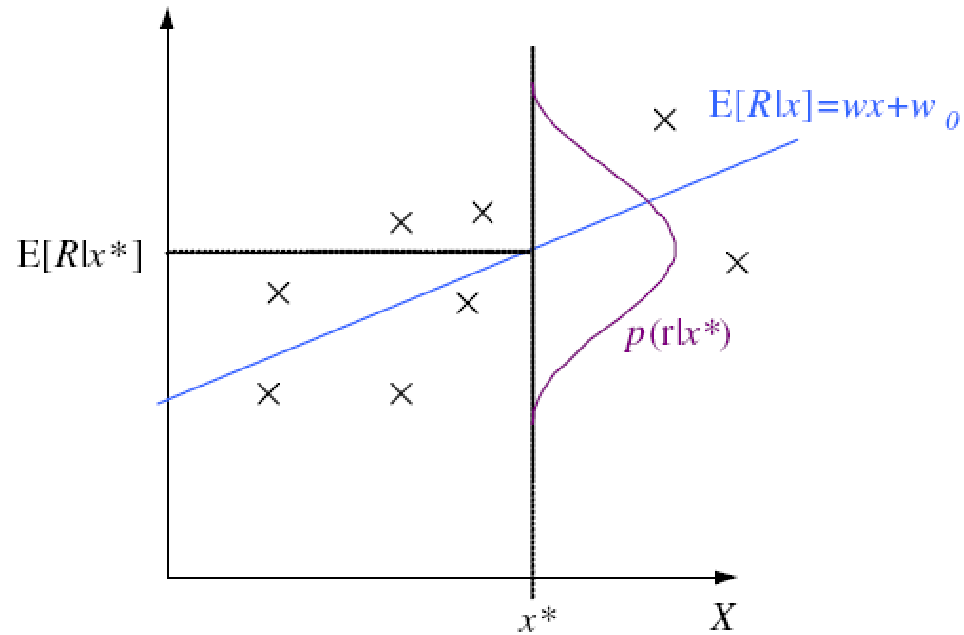
$$r = f(x) + \varepsilon$$

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

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

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

A value out of g() add to a Gaussian distribution -> a Gaussian distribution



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

Joint = condition x unconditioned

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

Likelihood function

Regression: From LogL to Error

Expected value

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

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

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

Least square estimation

Linear Regression

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

This is our loss function

$$\boxed{g(x^t \mid w_1, w_0) = w_1 x^t + w_0}$$

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

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

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

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

Polynomial Regression

$$g\left(x^t \mid w_k,\ldots,w_2,w_1,w_0\right)=w_k\left(x^t\right)^k+\cdots+w_2\left(x^t\right)^2+w_1x^t+w_0$$

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

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

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

If it is close to 1, it is almost like guessing the average of the data at all time

- Absolute Error: $E(\vartheta|\mathrm{X}) = \sum_t |r^t-g(x^t|\vartheta)|$

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

Error always decreases with more complex model
So how do we choose and select model?

Bias and Variance for regression

No g, just pure noise

Difference between real output and our estimate output
Quantify how well on average our g(x) is on the training set

$$E\left[(r-g(x))^2 \mid x\right] = \underbrace{\boxed{E\left[(r-E[r \mid x])^2 \mid x\right]}}_{noise} + \underbrace{\boxed{(E[r \mid x]-g(x))^2}}_{squared\ error}$$

$$E_X\left[(E[r \mid x]-g(x))^2 \mid x\right] = \underbrace{\boxed{(E[r \mid x]-E_X[g(x)])^2}}_{bias} + \underbrace{\boxed{E_X\left[(g(x)-E_X[g(x)])^2\right]}}_{variance}$$

To quantify how well on average our g(x) is on different dataset, we take the average over X

Side note:
Var(x) = E(X^2)-E(X)^2
Var(x) = E((x-E(x))^2)

Estimating Bias and Variance

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

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

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

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

Bias/Variance Dilemma

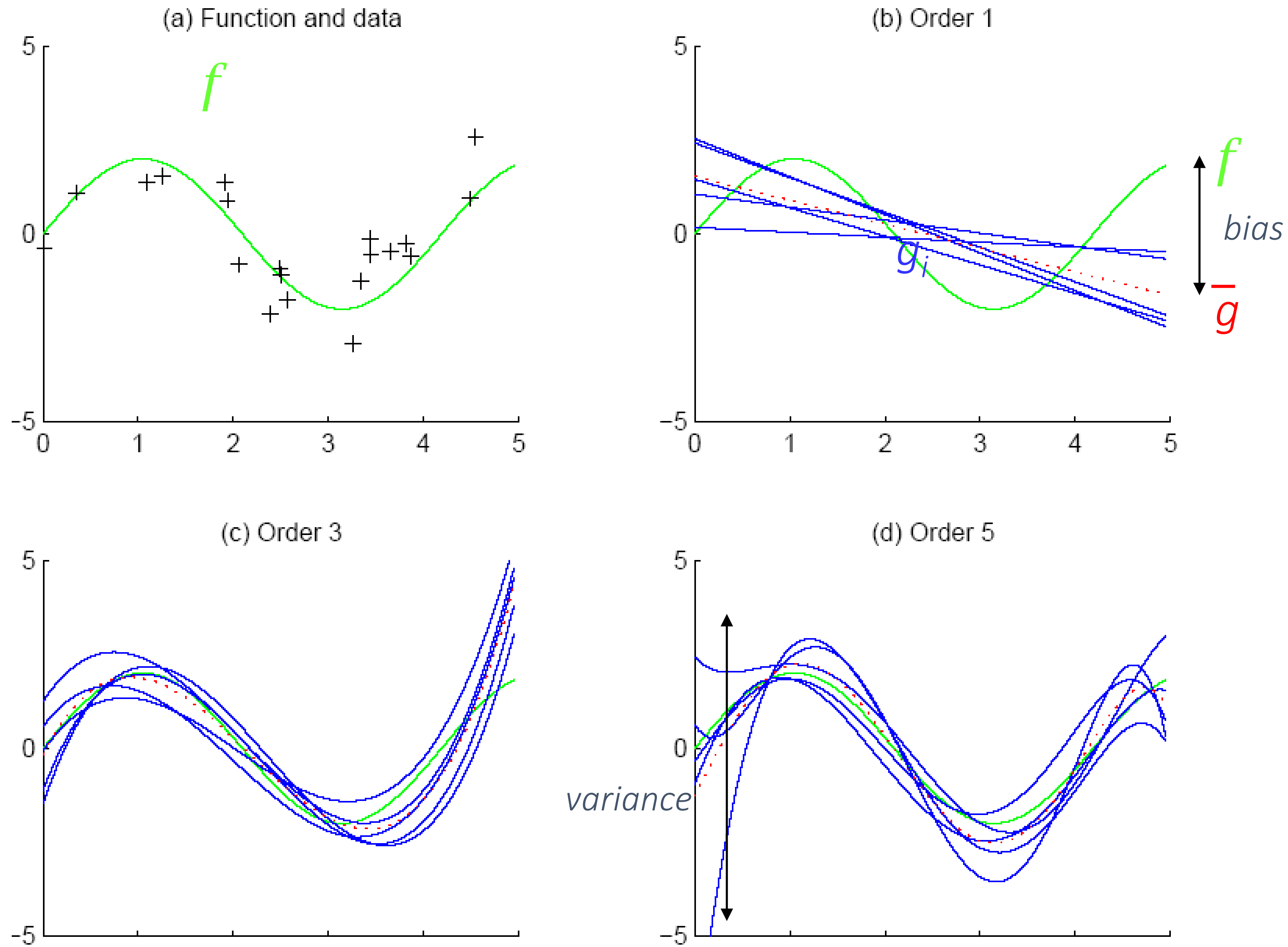
- Example: $g_i(x)=2$ has no variance and high bias

  $g_i(x)= \sum_t r^t_i/N$ 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
g
(c) Order 3
(d) Order 5
variance