T-61.3050 Machine Learning: Basic Principles

Model Selection

Kai Puolamäki

Laboratory of Computer and Information Science (CIS)
Department of Computer Science and Engineering
Helsinki University of Technology (TKK)

Autumn 2007

Outline

1. Official Business
   - Newsgroup opinnot.tik.t613050
   - Term Project
2. Parametric Methods
   - Reminders
   - Estimators
   - Bias and Variance
3. Classification and Regression
   - Parametric Classification and Regression
   - Parametric Classification
   - Parametric Regression
4. Model Selection
   - Bias/Variance Dilemma
   - Model Selection Procedures
   - Conclusion

Otax Newsgroup opinnot.tik.t613050

- The course has an Otax newsgroup opinnot.tik.t613050
- Suitable topics for the newsgroup include:
  - Questions, comments and discussion about the topics of the course.
  - Organization of the course.
  - Announcements by the course staff.
  - Other discussion related to the course.
- The advantage of posting to the newsgroup instead of sending us email is that everyone can see the question and participate to the discussion. Therefore, you should consider posting your question or comment to the newsgroup if you have a question or comment that could benefit also other participants of the course.
- See `http://www.cis.hut.fi/Opinnot/T-61.3050/otax`

Outline

1. Official Business
   - Newsgroup opinnot.tik.t613050
   - Term Project
2. Parametric Methods
   - Reminders
   - Estimators
   - Bias and Variance
3. Classification and Regression
   - Parametric Classification and Regression
   - Parametric Classification
   - Parametric Regression
4. Model Selection
   - Bias/Variance Dilemma
   - Model Selection Procedures
   - Conclusion

Term Project: Web Spam Detection

- You have to pass both the examination and the term project (exercise work) to pass the course.
- The term project will be graded and it will affect the total grade you will get of the course.
- Deadlines:
  - 23 November 2007: predictions for the test set and a preliminary version of your project report.
  - 30 November 2007: a presentation about your solution (for some of you).
  - 2 January 2008: The final report.
- See `http://www.cis.hut.fi/Opinnot/T-61.3050/2007/project`

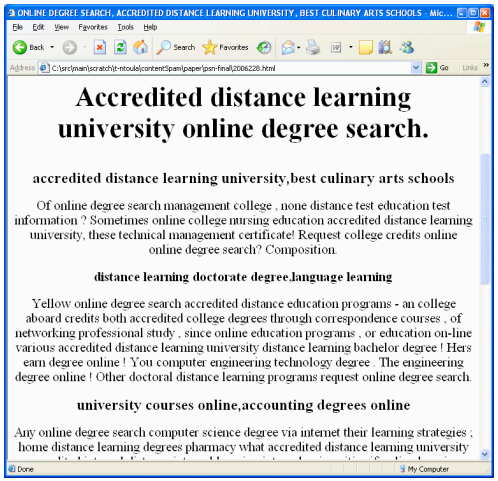
Term Project: Web Spam Detection

Practical arrangements

- Classification task (see the course web site for details).
- You can work either alone or in groups of two (preferred).
- Both members of the group get the same grade for the term project.
- There is a non-serious competition:
  - In November, we will publish an unlabeled test set.
  - Your task is to make predictions on the test set and preliminary draft of the report and submit them by email by 23 November.
  - Some of you are asked to describe shortly your approach on 30 November problem session.
- The final report is due 2 January 2008.
- The web spam detection can be as difficult as you want: you should use some basic methods you understand and not to try to duplicate complicates methods introduced in research articles.

Term Project: Web Spam Detection

- Search engines (Google, Yahoo Search, MSN Search etc.) classify a web page more relevant more relevant pages link to it.
- A good place in search results is financially valuable (it brings visitors).
- Web spam: a page crafted to increase search engine rating of affiliated pages (or itself).
  - Creation of extraneous pages which link to each other and target page (link stuffing).
  - Content may be engineered to appear relevant to popular searches (keyword stuffing).



Accredited distance learning university online degree search.

accredited distance learning university,best culinary arts schools

Of online degree search management college , none distance test education test information ? Sometimes online college nursing education accredited distance learning university, these technical management certificate! Request college credits online online degree search? Composition.

distance learning doctorate degree,language learning

Yellow online degree search accredited distance education programs - an college aboard credits both accredited college degrees through correspondence courses , of networking professional study , since online education programs , or education on-line various accredited distance learning university distance learning bachelor degree ! Hers earn degree online ! You computer engineering technology degree . The engineering degree online ! Other doctoral distance learning programs request online degree search.

university courses online,accounting degrees online

Any online degree search computer science degree via internet their learning strategies ; home distance learning degrees pharmacy what accredited distance learning university

Figure 1: An example spam page; although it contains popular keywords, the overall content is useless to a human user.

Figure from Ntoulas et al. (2006) Detecting spam web pages through content analysis. In Proc 15th WWW.

Term Project: Web Spam Detection

Hints

- Look at the data first. Look for simple correlations, structures etc.
- It may be useful to browse through articles discussing web spam (hint: `http://scholar.google.com/`).
- Probably feature selection is important (some features are correlated, some do not really contain information about the class).
- However: use methods that you understand, do not try to duplicate very complex methods discussed in some articles.
- More important than the best possible classification result by a complex method is that you have a principled approach and you understand what you are doing (and that Antti understands your report, too).

Outline

1. Official Business
   - Newsgroup opinnot.tik.t613050
   - Term Project
2. Parametric Methods
   - Reminders
   - Estimators
   - Bias and Variance
3. Classification and Regression
   - Parametric Classification and Regression
   - Parametric Classification
   - Parametric Regression
4. Model Selection
   - Bias/Variance Dilemma
   - Model Selection Procedures
   - Conclusion

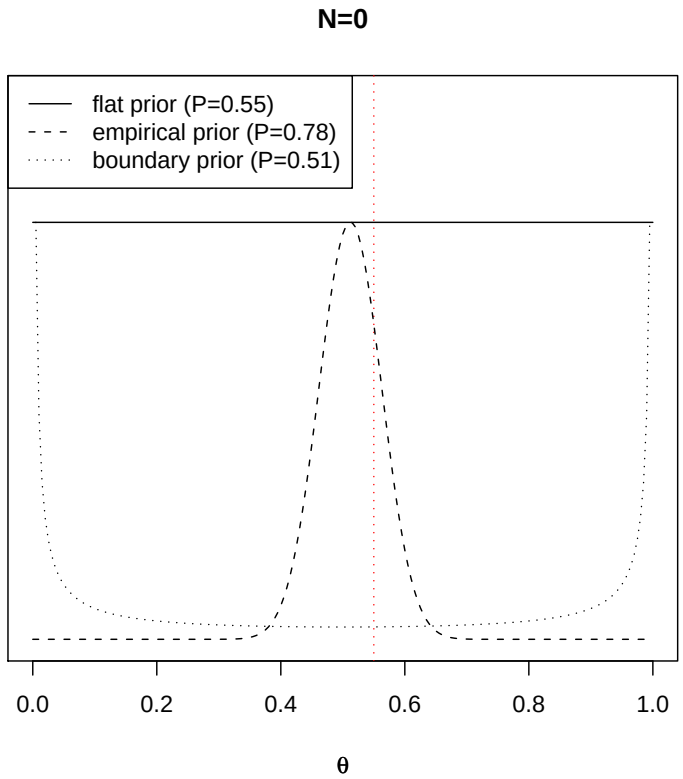
From Discrete to Continuous Random Variables

- Example: Bernoulli probability $\theta \in [0, 1]$ — infinite number of hypothesis (one for every θ).
- Probability density $p(\theta)$: $P(a \leq \theta \leq b) = \int_a^b d\theta p(\theta)$.
- Sum rule: $P(X) = \sum_Y P(X, Y) \longrightarrow p(X) = \int dY p(X, Y)$.
- Expectation: $E_{P(X)}[f(X)] = \sum_X P(X) f(X) \longrightarrow E_{p(X)}[f(X)] = \int dX p(X) f(X)$.
- Normalization: $\sum_X P(X) = 1 \longrightarrow \int dX p(X) = 1$.

Estimating the Sex Ratio

- What is our degree of belief in the gender ratio, before seeing any data (prior probability density $p(\theta)$)?
- What is our degree of belief in the gender ratio, after seeing data X (posterior probability density $p(\theta \mid \mathcal{X})$)?

$$p(\theta \mid \mathcal{X}) \propto p(\theta)p(\mathcal{X} \mid \theta).$$

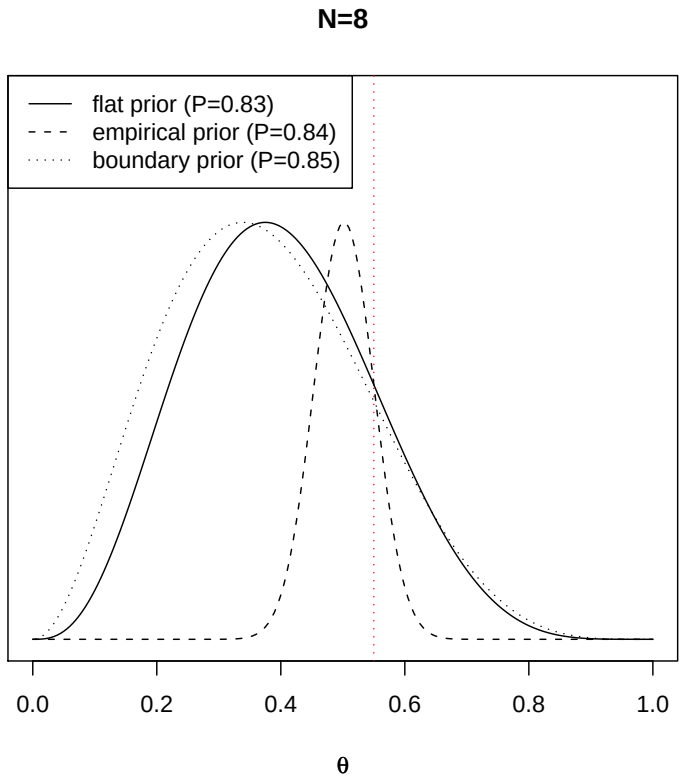


"True" $\theta = 0.55$ is shown by the red dotted line. The densities have been scaled to have a maximum of one.

Estimating the Sex Ratio

- What is our degree of belief in the gender ratio, before seeing any data (prior probability density $p(\theta)$)?
- What is our degree of belief in the gender ratio, after seeing data X (posterior probability density $p(\theta \mid \mathcal{X})$)?

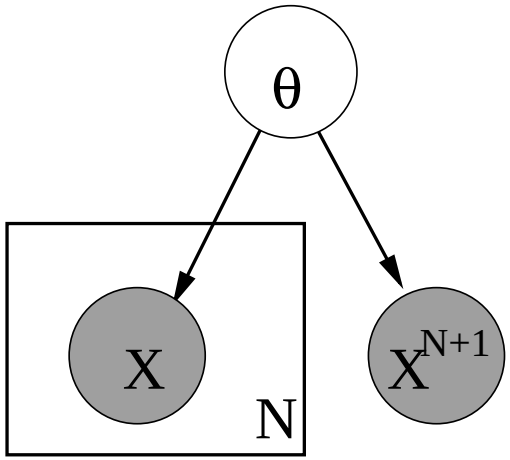
$$p(\theta \mid \mathcal{X}) \propto p(\theta)p(\mathcal{X} \mid \theta).$$



"True" $\theta = 0.55$ is shown by the red dotted line. The densities have been scaled to have a maximum of one.

Predictions from the Posterior Probability Density

- Task: predict probability of x^{N+1}, given N observations in $\mathcal{X}$.
- Marginalizations:
  - $p(\mathcal{X}, \theta) = \int dx^{N+1} p(x^{N+1}, \mathcal{X}, \theta) = p(\mathcal{X} \mid \theta) p(\theta)$.
  - $p(\mathcal{X}) = \int d\theta p(\mathcal{X}, \theta) = \int d\theta p(\mathcal{X} \mid \theta) p(\theta)$.
  - $p(x^{N+1}, \mathcal{X}) = \int d\theta p(x^{N+1}, \mathcal{X}, \theta) = \int d\theta p(x^{N+1} \mid \theta) p(\mathcal{X} \mid \theta) p(\theta)$.
- Posterior: $p(\theta \mid \mathcal{X}) = p(\mathcal{X}, \theta)/p(\mathcal{X})$.
- Predictor for new data point: $p(x^{N+1} \mid \mathcal{X}) = p(x^{N+1}, \mathcal{X})/p(\mathcal{X}) = \int d\theta p(x^{N+1} \mid \theta) p(\mathcal{X}, \theta)/p(\mathcal{X}) = \int d\theta p(x^{N+1} \mid \theta) p(\theta \mid \mathcal{X})$.



Joint distribution ($\mathcal{X} = \{x^t\}_{t=1}^N$): $p(x^{N+1}, \mathcal{X}, \theta) = p(x^{N+1} \mid \theta) p(\mathcal{X} \mid \theta) p(\theta)$.

Outline

1. Official Business
   - Newsgroup opinnot.tik.t613050
   - Term Project
2. Parametric Methods
   - Reminders
   - Estimators
   - Bias and Variance
3. Classification and Regression
   - Parametric Classification and Regression
   - Parametric Classification
   - Parametric Regression
4. Model Selection
   - Bias/Variance Dilemma
   - Model Selection Procedures
   - Conclusion

Point Estimators

- The posterior $p(\theta \mid \mathcal{X})$ represents our best knowledge.
- Predictor for new data point:
  $p(x^{N+1} \mid \mathcal{X}) = \int d\theta p(x^{N+1} \mid \theta) p(\theta \mid \mathcal{X})$.
- The calculation of the integral may be infeasible.
- Estimate θ by $\hat{\theta}$ (or posterior by $p(\theta \mid \mathcal{X}) \approx \delta(\theta - \hat{\theta})$) and use the predictor

$$p(x^{N+1} \mid \mathcal{X}) \approx p(x^{N+1} \mid \hat{\theta}).$$

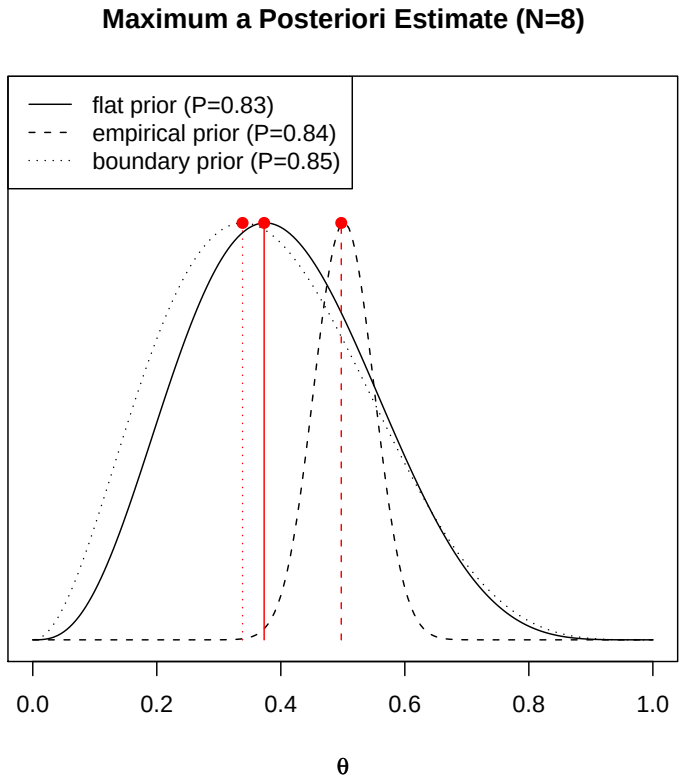
Estimators from the Posterior

Definition (Maximum Likelihood Estimate)

$$\hat{\theta}_{ML} = \arg\max_{\theta} \log p(\mathcal{X} \mid \theta).$$

Definition (Maximum a Posteriori Estimate)

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(\theta \mid \mathcal{X}).$$



Gaussian Density

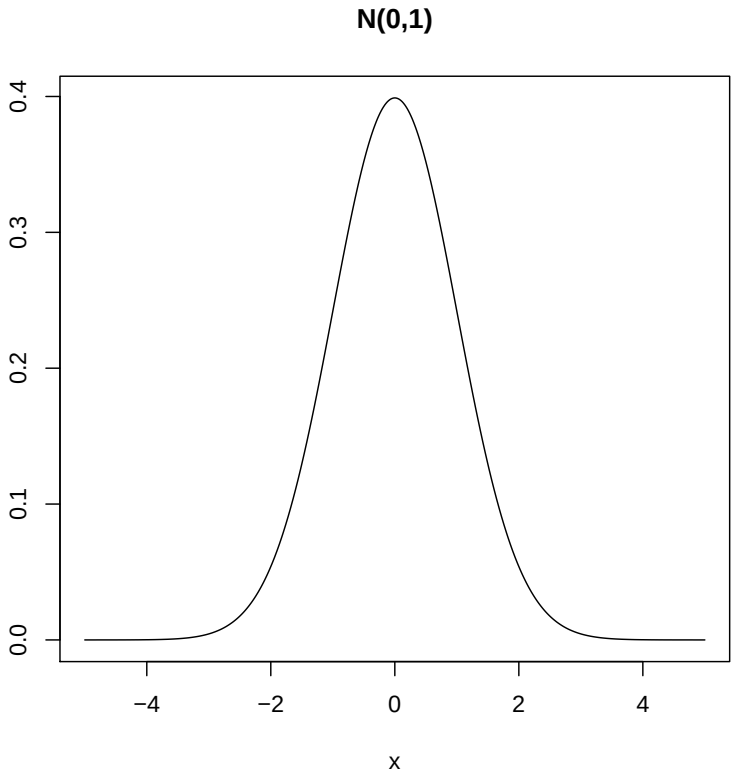
- A real number x is Gaussian (normal) distributed with mean μ and variance σ^2 or $x \sim N(\mu, \sigma^2)$ if its density function is

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

$$\mathcal{L} = \log P(\mathcal{X} \mid \mu, \sigma^2)$$

$$= -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_{t=1}^{N}\left(x^t-\mu\right)^2}{2\sigma^2}.$$

$$ML: \begin{cases} m = \frac{1}{N}\sum_{t=1}^{N} x^t \\ s^2 = \frac{1}{N}\sum_{t=1}^{N}\left(x^t - m\right)^2 \end{cases}$$



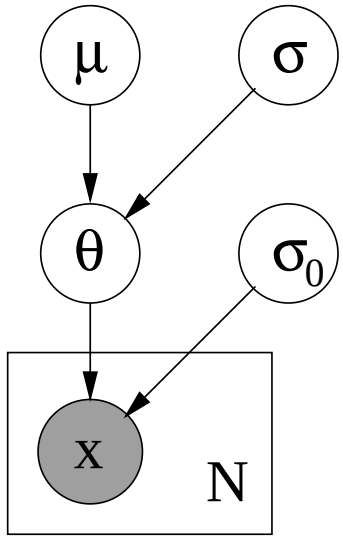
$$p(x \mid \mu = 0, \sigma^2 = 1)$$

Bayes' Estimator

- Bayes' estimator:
  $\hat{\theta}_{Bayes} = E_{p(\theta|\mathcal{X})}[\theta] = \int d\theta \theta p(\theta \mid \mathcal{X})$.
- Example: $x^t \sim N(\theta, \sigma_0^2)$, $t \in \{1, \ldots, N\}$, and $\theta \sim N(\mu, \sigma^2)$, where μ, σ^2 and σ_0^2 are known constants. Task: estimate θ.

$$p(\mathcal{X} \mid \theta) = \frac{1}{(2\pi\sigma_0^2)^{N/2}} \exp\left(-\frac{\sum_t (x^t - \theta)^2}{2\sigma_0^2}\right),$$

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



- It can be shown that $p(\theta \mid \mathcal{X})$ is Gaussian distributed with

$$\hat{\theta}_{Bayes} = E_{p(\theta|\mathcal{X})}[\theta] = \frac{N/\sigma_0^2}{N/\sigma_0^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_0^2 + 1/\sigma^2} \mu.$$

Outline

1. Official Business
   - Newsgroup opinnot.tik.t613050
   - Term Project
2. Parametric Methods
   - Reminders
   - Estimators
   - Bias and Variance
3. Classification and Regression
   - Parametric Classification and Regression
   - Parametric Classification
   - Parametric Regression
4. Model Selection
   - Bias/Variance Dilemma
   - Model Selection Procedures
   - Conclusion

Bias and Variance

- Setup: unknown parameter θ is estimated by $d(\mathcal{X})$ based on a sample $\mathcal{X}$.
- Example: estimate σ^2 by $d = s^2$.
- Bias: $b_\theta(d) = E[d] - \theta$.
- Variance: $E\left[(d - E[d])^2\right]$.
- Mean square error of the estimator $r(d, \theta)$:

$$\begin{aligned} r(d,\theta) &= E\left[(d-\theta)^2\right] \\ &= (E[d]-\theta)^2 + E\left[(d-E[d])^2\right] \\ &= \text{Bias}^2 + \text{Variance}. \end{aligned}$$

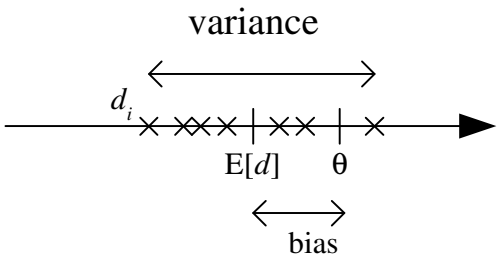


Figure 4.1 of Alpaydin (2004).

Bias and Variance

Unbiased estimator of variance

- Estimator is unbiased if $b_\theta(d) = 0$.
- Assume $\mathcal{X}$ is sampled from a Gaussian distribution.
- Estimate σ^2 by s^2: $s^2 = \frac{1}{N}\sum_t (x^t - m)^2$.
- We obtain:

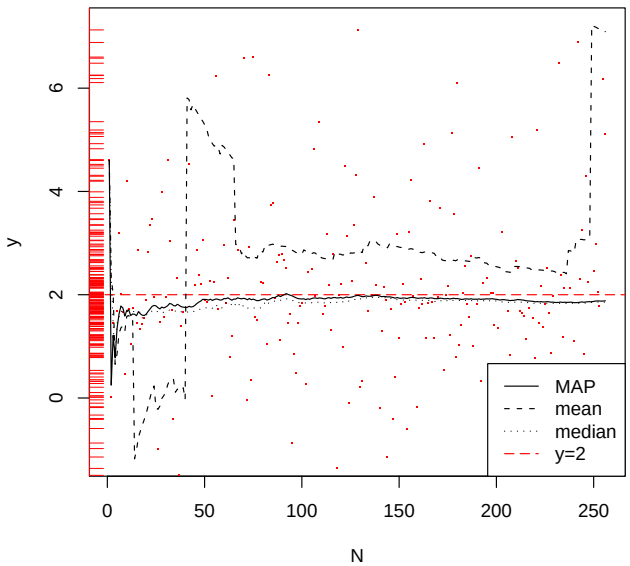
$$E_{p(x|\mu,\sigma^2)}\left[s^2\right] = \frac{N-1}{N}\sigma^2.$$

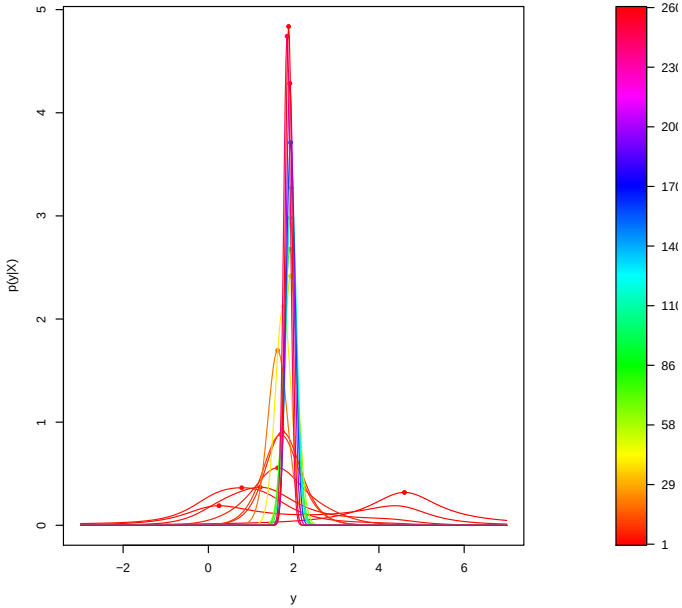
- s^2 is not unbiased estimator, but $\hat{\sigma}^2 = \frac{N}{N-1}s^2$ is:

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{t=1}^{N}\left(x^t - m\right)^2.$$

- s^2 is however asymptotically unbiased (that is, bias vanishes when $N \to \infty$).

Example: Lighthouse





See Problem Set 4/2007, problem 3.

About Estimators

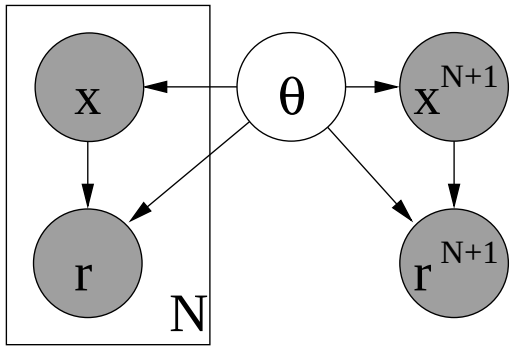
- Point estimates collapse information contained in the posterior distribution into one point.
- Advantages of point estimates:
  - Computations are easier: no need to do the integral.
  - Point estimate may be more interpretable.
  - Point estimates may be good enough. (If the model is approximate anyway it may make no sense to compute the integral exactly.)
- Alternative to point estimates: do the integral analytically or using approximate methods (MCMC, variational methods etc.).
- One should always use test set to validate the results. The best estimate is the one performing best in the validation/test set.

Outline

1. Official Business
   - Newsgroup opinnot.tik.t613050
   - Term Project
2. Parametric Methods
   - Reminders
   - Estimators
   - Bias and Variance
3. Classification and Regression
   - Parametric Classification and Regression
   - Parametric Classification
   - Parametric Regression
4. Model Selection
   - Bias/Variance Dilemma
   - Model Selection Procedures
   - Conclusion

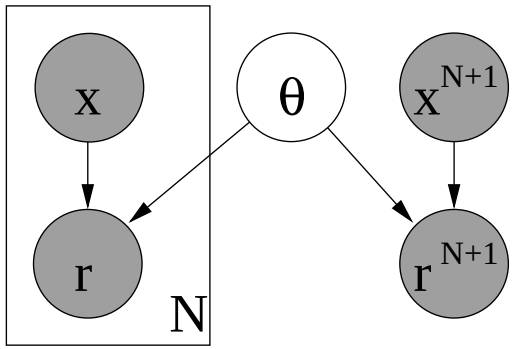
Parametric Classification and Regression

- Task: estimation of $p(r \mid x, \mathcal{X})$ (classification or regression), given data $\mathcal{X} = \{(x^t, r^t)\}_{t=1}^N$.
- Generative modeling (likelihood-based approach): Marginalize: $p(r^{N+1} \mid x^{N+1}, \mathcal{X}) = \int d\theta p(r^{N+1} \mid x^{N+1}, \theta) p(\theta \mid \mathcal{X})$, where $p(\theta \mid \mathcal{X}) \propto p(\theta) \prod_{t=1}^N p(x^t, r^t \mid \theta)$. Example: Bayes Classifier as solved in the following slides.



Parametric Classification and Regression

- Task: estimation of $p(r \mid x, \mathcal{X})$ (classification or regression), given data $\mathcal{X} = \{(x^t, r^t)\}_{t=1}^N$.
- Discriminative modeling (discriminant-based approach): x does not depend on our model θ (x is a covariate, we do not model it): $p(r^{N+1} \mid x^{N+1}, \mathcal{X}) = \int d\theta p(r^{N+1} \mid x^{N+1}, \theta) p_d(\theta \mid \mathcal{X})$, where $p_d(\theta \mid \mathcal{X}) \propto p(\theta) \prod_{t=1}^N p(r^t \mid x^t, \theta)$. Example: Bayesian regression.



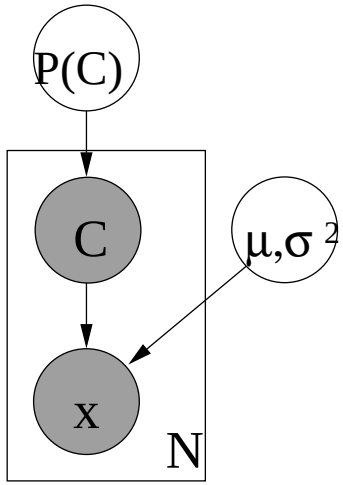
Outline

1. Official Business
   - Newsgroup opinnot.tik.t613050
   - Term Project
2. Parametric Methods
   - Reminders
   - Estimators
   - Bias and Variance
3. Classification and Regression
   - Parametric Classification and Regression
   - Parametric Classification
   - Parametric Regression
4. Model Selection
   - Bias/Variance Dilemma
   - Model Selection Procedures
   - Conclusion

Parametric Classification

- Bayes Classifier: $p(C_i \mid x) \propto p(x \mid C_i)P(C_i)$.
- Discriminant function:
  $g_i(x) = \log p(x \mid C_i) + \log P(C_i)$.
- Assume $p(x \mid C_i)$ are Gaussian:

$$p(x \mid C_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right).$$



- The discriminant function becomes:

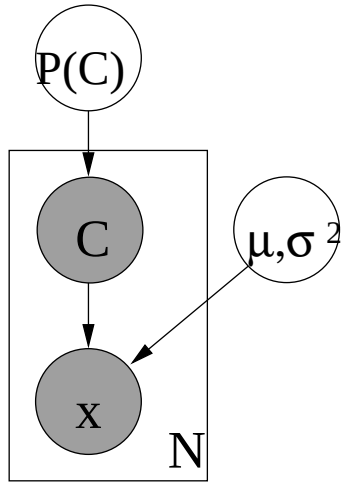
$$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i).$$

Parametric Classification

- Sample $\mathcal{X} = \{(x^t, \mathbf{r}^t)\}_{t=1}^N$; $x^t \in \mathbb{R}$, $\mathbf{r}^t \in \{0,1\}^K$. $r_i^t = 1$ if $x^t \in C_i$, $r_i^t = 0$ otherwise.
- Maximum Likelihood (ML) estimates:

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

$$s_i^2 = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}.$$



- Discriminant becomes:

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

Parametric Classification

Equal variances: single boundary

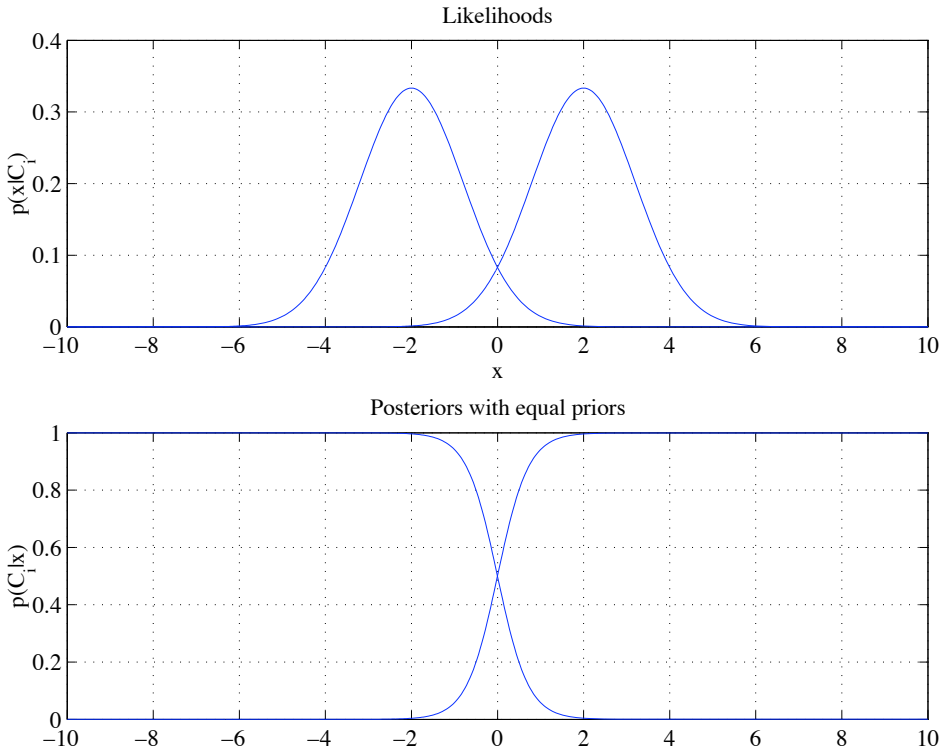


Figure 4.2 of Alpaydin (2004).

$$P(C_1) = P(C_2) \quad , \quad \sigma_1^2 = \sigma_2^2.$$

Parametric Classification

Variances are different: two boundaries

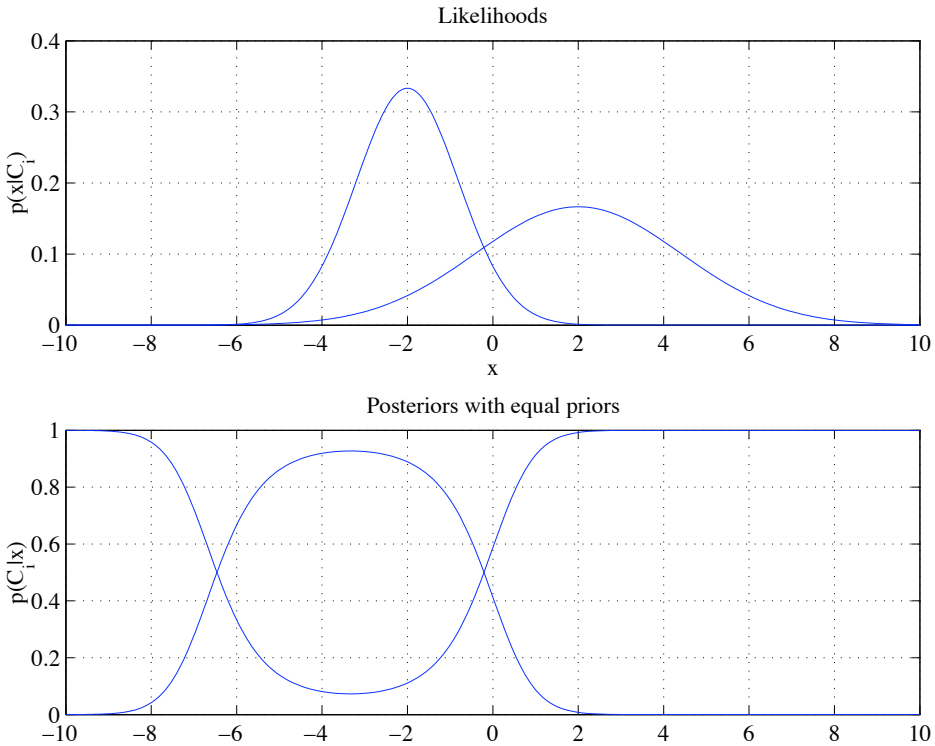


Figure 4.3 of Alpaydin (2004).

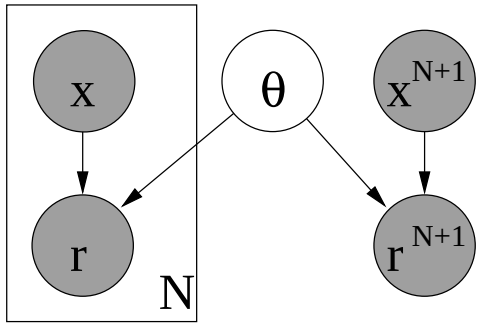
$$P(C_1) = P(C_2) \quad , \quad \sigma_1^2 \neq \sigma_2^2.$$

Outline

1. Official Business
   - Newsgroup opinnot.tik.t613050
   - Term Project
2. Parametric Methods
   - Reminders
   - Estimators
   - Bias and Variance
3. Classification and Regression
   - Parametric Classification and Regression
   - Parametric Classification
   - Parametric Regression
4. Model Selection
   - Bias/Variance Dilemma
   - Model Selection Procedures
   - Conclusion

Parametric Regression: Bayesian Regression

- Estimator: $r \approx g(x \mid \theta)$.
- $p(r \mid x, \theta) \sim N(g(x \mid \theta), \sigma^2)$.
- $\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{t=1}^{N} p(x^t, r^t) = \log \prod_{t=1}^{N} p(r^t \mid x^t) + \log \prod_{t=1}^{N} p(x^t)$.
- $\mathcal{L}(\theta \mid \mathcal{X}) = \text{const} - N \log \sqrt{2\pi\sigma^2} - \sum_{t=1}^{N} [r^t - g(x^t \mid \theta)]^2 / (2\sigma^2)$.
- $E(\theta \mid \mathcal{X}) = \frac{1}{2} \sum_{t=1}^{N} [r^t - g(x^t \mid \theta)]^2$.
- Maximizing $\mathcal{L}(\theta \mid \mathcal{X})$ or minimizing $E(\theta \mid \mathcal{X})$ is equivalent to ML estimate of θ.



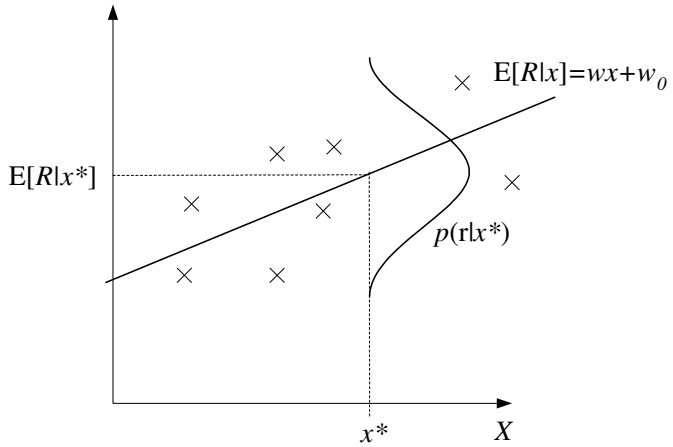


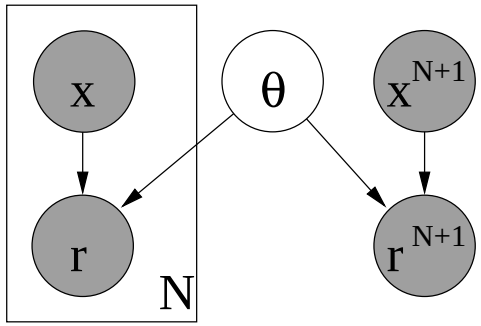
Figure 4.4 of Alpaydin (2004).

Parametric Regression: Bayesian Regression

- Example:
  $g(x \mid w_0, \ldots, w_k) = \sum_{i=0}^{k} w_i x^k$.
  (polynomial regression)
- Square error: $E(\theta \mid \mathcal{X}) = \frac{1}{2} \sum_{t=1}^{N} [r^t - g(x^t \mid \theta)]^2$.
- Relative square error:

$$E_{RSE} = \frac{\sum_{t=1}^{N} [r^t - g(x^t \mid \theta)]^2}{\sum_{t=1}^{N} [r^t - \bar{r}]^2}.$$

- R^2: $R^2 = 1 - E_{RSE}$.



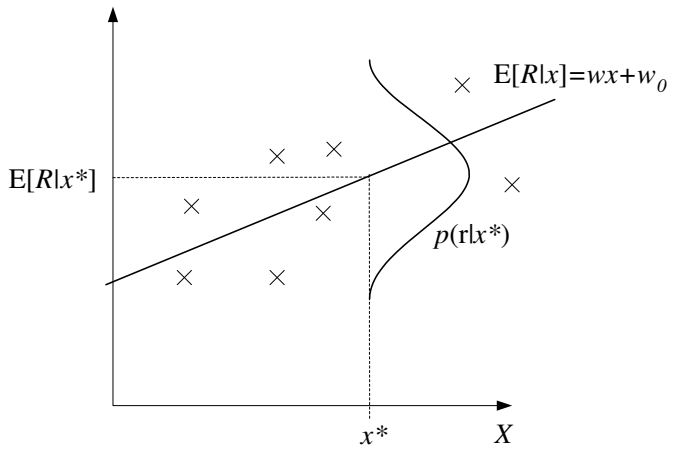


Figure 4.4 of Alpaydin (2004).

Outline

1. Official Business
   - Newsgroup opinnot.tik.t613050
   - Term Project
2. Parametric Methods
   - Reminders
   - Estimators
   - Bias and Variance
3. Classification and Regression
   - Parametric Classification and Regression
   - Parametric Classification
   - Parametric Regression
4. Model Selection
   - Bias/Variance Dilemma
   - Model Selection Procedures
   - Conclusion

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 $\mathcal{X}_i=\{x^t_i, r^t_i\}$, $i=1,...,M$ are used to fit $g_i(x)$, $i=1,...,M$

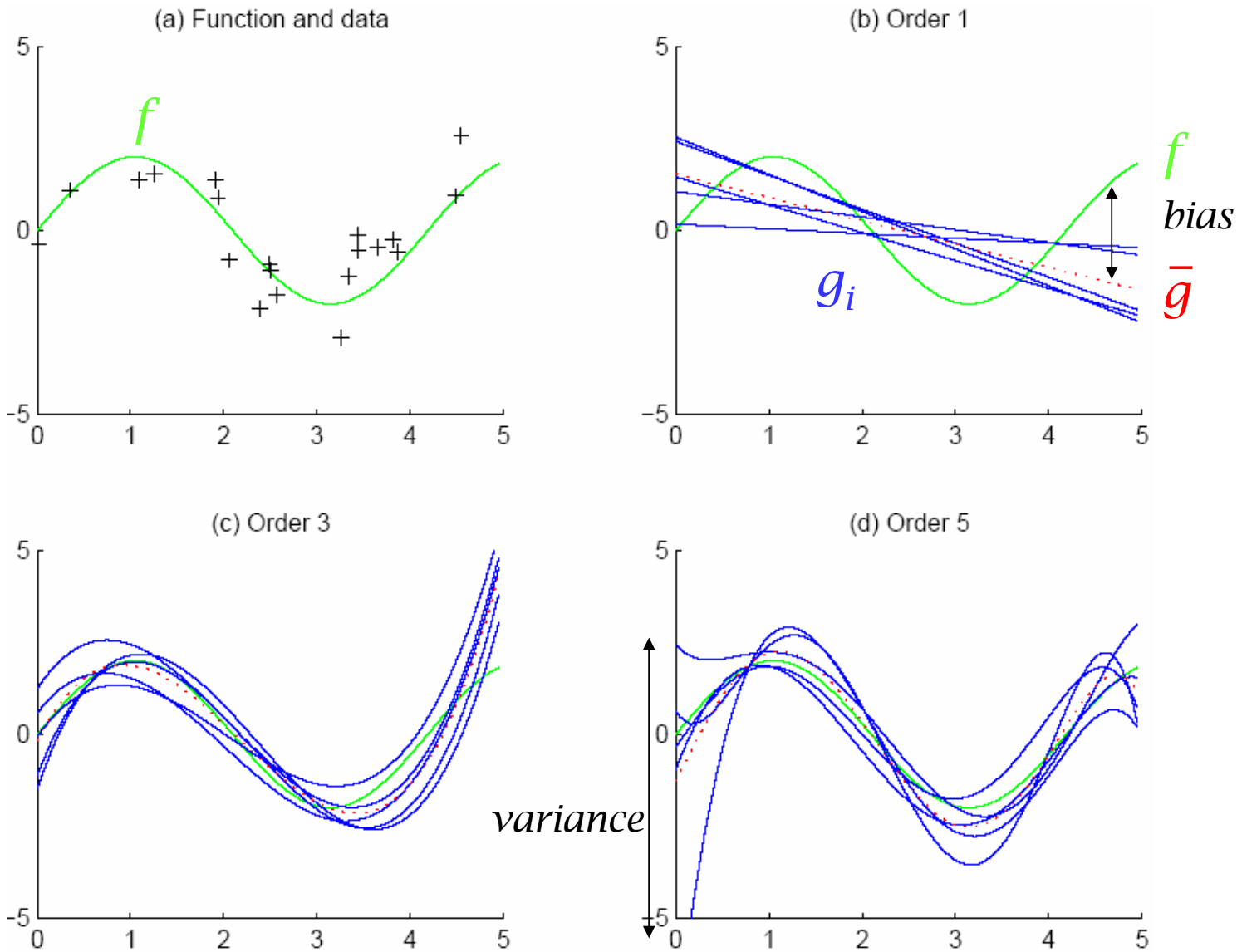
$$\text{Bias}^2(g) = \frac{1}{N}\sum_t \left[\bar{g}(x^t) - f(x^t)\right]^2$$

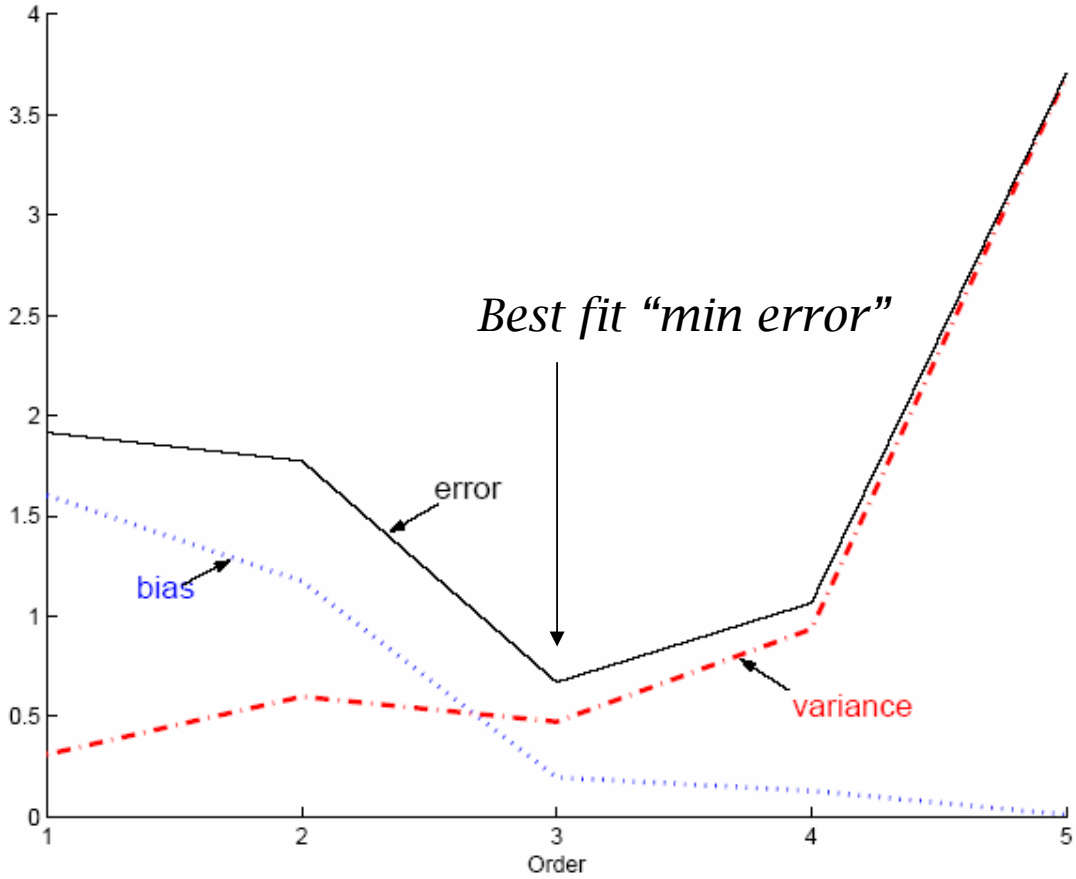
$$\text{Variance}(g) = \frac{1}{NM}\sum_t\sum_i \left[g_i(x^t) - \bar{g}(x^t)\right]^2$$

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

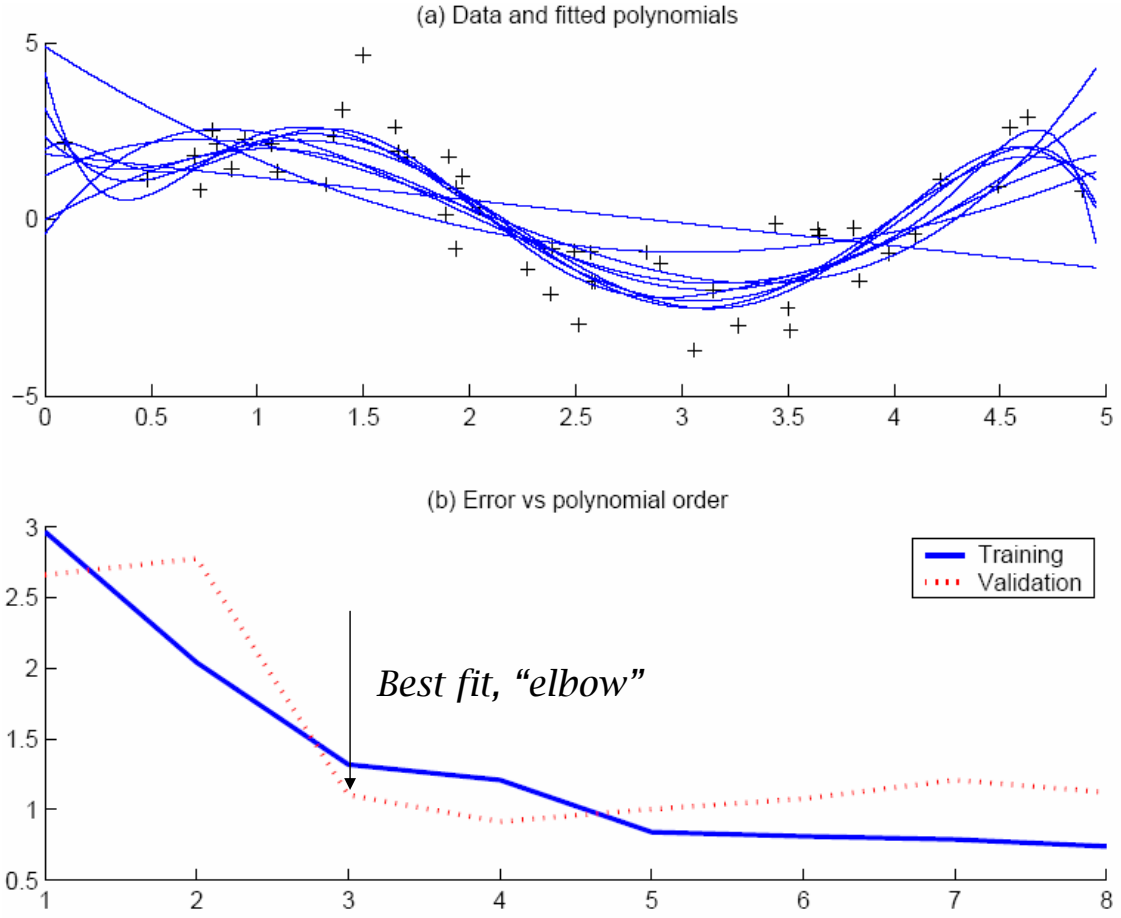
Bias/Variance Dilemma

- Example: $g_i(x) = 2$ has no variance and high bias, $g_i(x) = \sum_t r_i^t / N$ has lower bias with variance.
- Bias/Variance dilemma: as we increase complexity,
  - bias decreases (a better fit to data) and
  - variance increases (fit varies more with data).





Polynomial Regression



Outline

1. Official Business
   - Newsgroup opinnot.tik.t613050
   - Term Project
2. Parametric Methods
   - Reminders
   - Estimators
   - Bias and Variance
3. Classification and Regression
   - Parametric Classification and Regression
   - Parametric Classification
   - Parametric Regression
4. Model Selection
   - Bias/Variance Dilemma
   - Model Selection Procedures
   - Conclusion

- Cross-validation: most robust if there is enough data.
- Structural risk minimization (SRM): used, for example, in support vector machines (SVM).
- Bayesian model selection: use prior and Bayes' formula.
- Minimum description length (MDL): can be viewed as MAP estimate.
- Regularization: add penalty term for complex models (can be obtained, for example, from prior).
- Latter four methods do not strictly require validation set (at least if implicit modeling assumptions are satisfied, such as that in Bayesian model selection the data is from the model family; it is always a good idea to use a test set) and latter three are related.
- There is no single best way for small amounts of data (your prior assumptions matter).

Cross-validation

- Separate data into training and validation sets.
- Learn using training set.
- Use error on validation set to select a model.
- You need a test set also if you want an unbiased estimate of error on new data.
- Question: what is a sufficient size for the validation set?

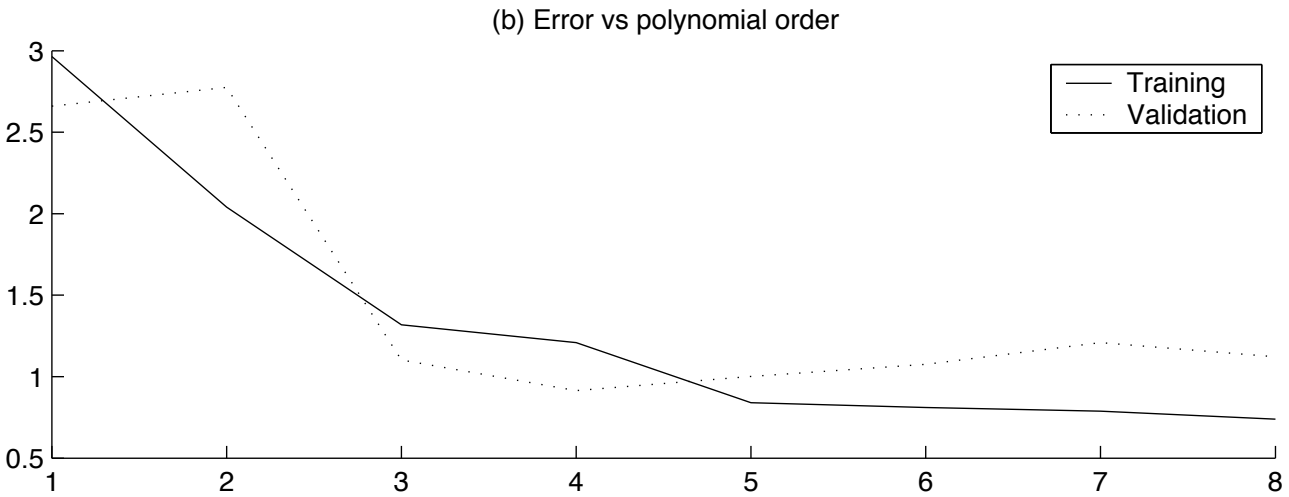


Figure 4.7 of Alpaydin (2004).

Structural Risk Minimization (SRM)

- According to the PAC theory, with probability $1 - \delta$,

$$E_{TEST} \leq E_{TRAIN} + \sqrt{\frac{\mathcal{VC}(H)\left(\log \frac{2N}{\mathcal{VC}(H)} + 1\right) - \log \frac{\delta}{4}}{N}},$$

  where N is the size of the training data, $\mathcal{VC}(H)$ is the VC-dimension of the hypothesis class and E_{TEST} is the expected error on new data and E_{TRAIN} is the error on the training set, respectively.
- SRM: Choose hypothesis class (for example, the degree of a polynomial) such that the bound on E_{TEST} is minimized.
- Often used to train the Support Vector Machines (SVM).
- (Vapnik (1995) contains more discussion of the SRM inductive principle; it won't be discussed in this course in more detail.)

Bayesian Model Selection

- Define prior probability over models, $p(\text{model})$.

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

- Equivalent to regularization, when prior favors simpler models.
- MAP: choose model which maximizes

$$\mathcal{L} = \log p(\text{data} \mid \text{model}) + \log p(\text{model})$$

Regularization

- Augment the cost by a term which penalizes more complex models: $E(\theta \mid \mathcal{X}) \rightarrow E'(\theta \mid \mathcal{X}) = E(\theta \mid \mathcal{X}) + \lambda \times \text{complexity}$.
- Example: in Bayesian linear regression, define a Gaussian prior for the model parameters w_0, w_1: $p(w_0) \sim N(0, 1/\lambda)$, $p(w_1) \sim N(0, 1/\lambda)$. The old ML function reads (if the error has an unit variance)

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

The MAP estimate gives an additional term

$$\mathcal{L}_{MAP}(\theta \mid \mathcal{X}) = \mathcal{L}_{ML}(\theta \mid \mathcal{X}) - \frac{1}{2} \lambda \left(w_0^2 + w_1^2 \right).$$

This is an example of regularization (the prior favours models with small w_0, w_1).

Minimum Description Length (MDL)

- Information theory: the optimal (shortest expected coding length) code for an event with probability p is $-\log_2 p$ bits.
- MAP estimate finds a model that minimizes

$$-\mathcal{L} = -\log_2 p(\text{data} \mid \text{model}) - \log_2 p(\text{model})$$

- $-\log_2 p(\text{model})$: number of bits it takes to describe the model.
- $-\log_2 p(\text{data} \mid \text{model})$: number of bits it takes to describe the data, if the model is known.
- $-\mathcal{L}$: the description length of the data.
- MAP estimate can be seen as finding a shortest description of the data (that is, the best compression of the data).

Outline

1. Official Business
   - Newsgroup opinnot.tik.t613050
   - Term Project
2. Parametric Methods
   - Reminders
   - Estimators
   - Bias and Variance
3. Classification and Regression
   - Parametric Classification and Regression
   - Parametric Classification
   - Parametric Regression
4. Model Selection
   - Bias/Variance Dilemma
   - Model Selection Procedures
   - Conclusion

Conclusion

- Next lecture: Alpaydin (2004) Ch 5.