浙江大学
ZHEJIANG UNIVERSITY

Mathematics in computer science

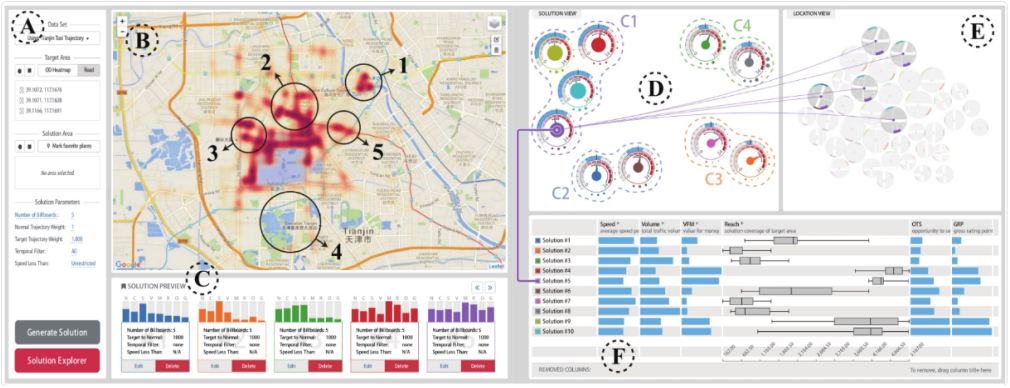
1. Point Estimation

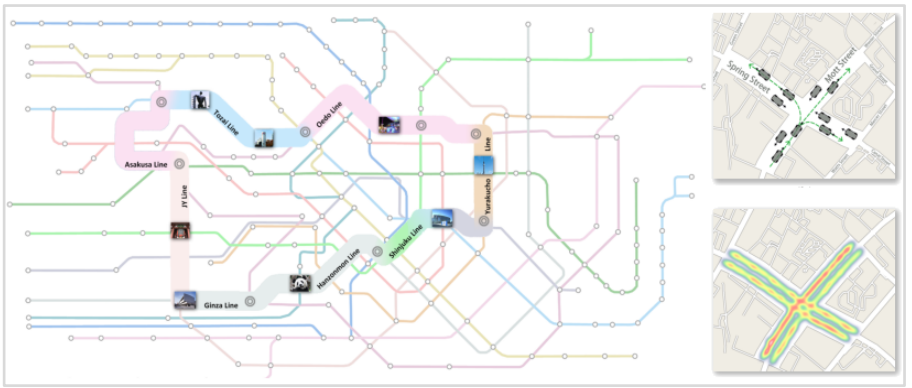
张宏鑫（Hongxin Zhang）

zhx@cad.zju.edu.cn

State Key Lab of CAD&CG, ZJU

2025-02-18



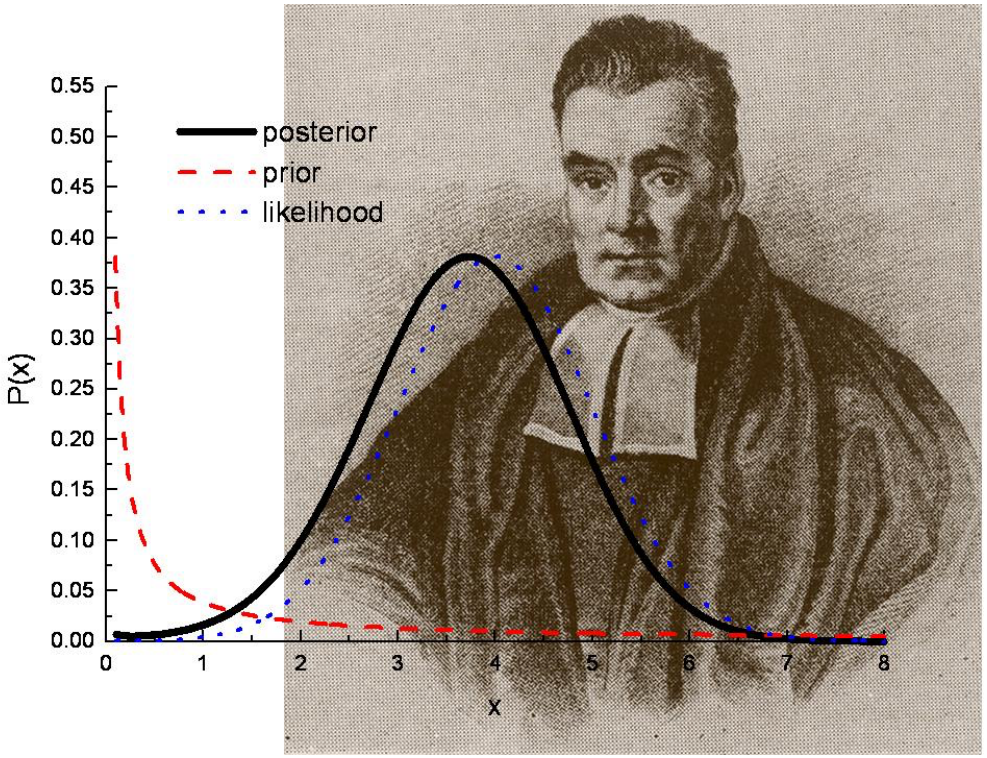


What you need to know

- Point estimation: (点估计)
  - **M**aximal **L**ikelihood **E**stimation (MLE)
  - Bayesian learning
  - **M**aximize **A** **P**osterior (MAP)
- Gaussian estimation
- Regression （回归）
  - Basis function = features
  - Optimizing sum squared error
  - Relationship between regression and Gaussians
- Bias-Variance trade-off

LLM Prompt (please try)

- “What is / Why / How to” +
  - Point estimation: (点估计)
    - **M**aximal **L**ikelihood **E**stimation (MLE)
    - Bayesian learning
    - **M**aximize **A** **P**osterior (MAP)
  - Gaussian estimation
  - Regression （回归）
    - Basis function = features
    - Optimizing sum squared error
    - Relationship between regression and Gaussians
  - Bias-Variance trade-off



Point estimation

点估计是非常重要的数据计算技术，需要在入门阶段重点掌握

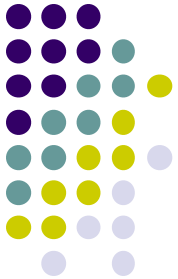
中国概率统计领域的发展

Your first consulting job

- An IT billionaire from USA asks you a question:
  - B: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
  - Y: Please flip it a few times …

  - Y: The probability is 3/5
  - B: Why???
  - Y: Because…

Binomial Distribution



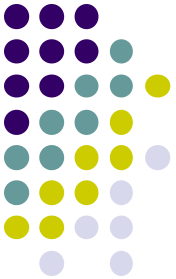
- $P(\text{Heads}) = \theta$, $P(\text{Tails}) = 1\text{-}\theta$ $\quad D = \{T, H, H, T, T\}$

$$P(D \mid \theta) = (1-\theta)\theta\theta(1-\theta)(1-\theta)$$

- Flips are i.i.d. (Independent Identically distributed)
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence D of α_H Heads and α_T Tails

$$P(D \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Maximum Likelihood Estimation



- **Data**: Observed set D of α_H Heads and α_T Tails
- **Hypothesis**: Binomial distribution
- **Learning θ is an optimization problem**
  - What's the objective function?

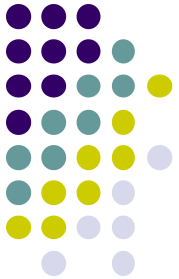
$$D = \{T, H, H, T, T\}$$

- **MLE**: Choose θ that maximizes the probability of observed data:

$$\hat{\theta} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \ln P(D \mid \theta) = \ldots$$

Maximum Likelihood Estimation (cont.)



$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} P(D|\theta) \\ &= \arg\max_{\theta} \ln(\theta^{\alpha_H}(1-\theta)^{\alpha_T}) \\ &= \arg\max_{\theta}(\alpha_H \ln\theta + \alpha_T \ln(1-\theta))\end{aligned}$$

- Set derivative to zero:

$$\frac{d}{d\theta}\ln P(D|\theta) = 0$$

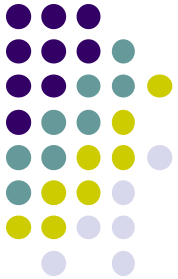
$$\hat{\theta} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{2}{2+3}$$

How many flips do I need?

$$\hat{\theta} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- B: I flipped 2 heads and 3 tails.
- Y: 1- θ = 3/5, I can prove it!
- B: What if I flipped 20 heads and 30 tails?
- Y: Same answer, I can prove it!
- B: What's better?
- Y: Humm… The more the merrier???
- B: Is this why I am paying you the big bucks???

Simple bound (based on Höffding's inequality)



- For $N = \alpha_H + \alpha_T$ and $\hat{\theta} = \dfrac{\alpha_T}{\alpha_H + \alpha_T}$

http://omega.albany.edu:8008/machine-learning-dir/notes-dir/vc1/vc-l.html

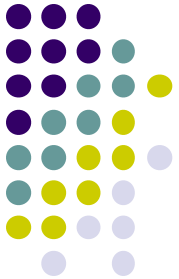
- Let θ* be the true parameter, for any ε>0:

$$P(\left|\hat{\theta} - \theta^*\right| \geq \varepsilon) \leq 2e^{-2N\varepsilon^2} \leq \delta$$

$$N \geq \frac{1}{2\varepsilon^2}[\ln 2 - \ln \delta]$$

$$N \geq 270; (\varepsilon = 0.1, \delta = 0.01)$$

PAC Learning



- **PAC**: **P**robably **A**pproximate **C**orrect
- B: I want to know the thumbtack parameter θ, within ε = 0.1, with probability at least 1-δ = 0.99. How many flips?
- Y: 270, ☺

Interval Estimation Basics

- **Interval Estimation:**
  Provides a likely range for an unknown population parameter
- **Binomial Focus:**
  Estimating the success probability (P) with confidence
- **Confidence Interval:**
  Data-based range reflecting estimation uncertainty
- **Advantage:**
  Offers a broader view than a single point estimate

Binomial Confidence Interval in Action

- **Problem:** Estimate the true heads probability (p) of a biased coin flipped 100 times with 60 heads observed
- **Point Estimate:** Sample proportion $\hat{p} = \frac{60}{100} = 0.6.$
- **95% Confidence Interval:**

  Using normal approximation: $X \sim N(np, np(1-p))$, $E(X) = np$, $\sigma = np(1-p)$
  $(n \geq 100), (np \geq 5), (n(1-p) \geq 5)$

  $$CI = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  $$= 0.6 \pm 1.96 \times \sqrt{\frac{0.6 \times 0.4}{100}}$$

  $$= 0.6 \pm 0.096$$

  $$= [0.504, 0.696]$$

- **Interpretation:** We are 95% confident that the true heads probability lies between 0.504 and 0.696. This interval quantifies the uncertainty around our point estimate, providing a more comprehensive understanding of p's possible values.
- **REF：** https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

Point ~ vs. Interval Estimation

Point Estimation:

- Specific value for unknown parameter.
- Binomial: often uses sample proportion (k/n).
- Simple, but no precision or reliability measure.

Interval Estimation:

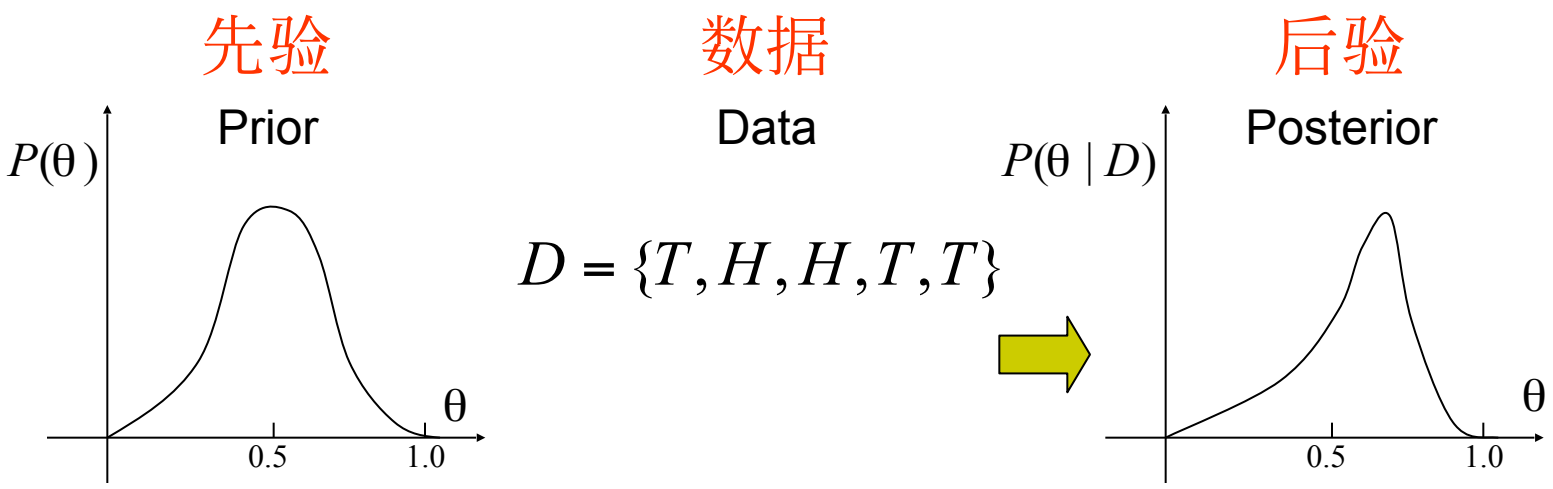
- Range of likely true parameter values.
- Binomial: confidence interval for p.
- Quantifies uncertainty, allows statistical inference.
- Narrower intervals = greater precision.

- **Point and Interval ~**
  - Point and interval estimation complement each other
  - Point estimate (e.g., sample proportion) is the starting point
  - Binomial confidence intervals:
    - normal approximation, Clopper-Pearson, etc
  - Reporting both gives precision and uncertainty
  - Essential for informed decisions and valid conclusions

Prior: knowledge before experiments

- B: Wait, I know that the thumbtack is “close” to 50-50. What can you …?
- Y: I can learn it the Bayesian way…

- Rather than estimating a single θ, we obtain a distribution over possible values of θ



从一个假想案例说起

- 王某去医院作验血化验，检查他患上了X疾病的可能性，其结果居然为阳性，把他吓了一大跳，赶忙到网上查询…

- 相关资料表明：该化验有
  - 1%假阳性率（False Positive）
    在得病的人中做实验，有1%的人是假阳性，99%的人是真阳性
  - 1%假阴性率（False Negative）
    在未得病的人中做实验，有1%的人是假阴性，99%的人是真阴性
- 问王某是否病了？或者说王某得病概率是否很大?

- 原始内容来源：http://news.sciencenet.cn/news/sub26.aspx?id=2958

从一个假想案例说起

不要慌 木有事

- 王某去医院作验血化验，检查他患上了X疾病的可能性，其结果居然为阳性，把他吓了一大跳，赶忙到网上查询…
- 相关资料表明：该化验有
  - 1%假阳性率
    在得病的人中做实验，有1%的人是假阳性，99%的人是真阳性
  - 1%假阴性率
    在未得病的人中做实验有1%的人是假阴性，99%的人是真阴性
- 关于王某得病概率，医生的回答是：9% （不要慌）
- 为什么：
  - 这种X疾病的正常比例是不大的，1000个人中只有一个人有X病

从一个假想案例说起

- 已知条件：
  - 相关资料表明：该化验有
    - 1%假阳性率
      在得病的人中做实验，有1%的人是假阳性，99%的人是真阳性
    - 1%假阴性率
      在未得病的人中实验，有1%的人是假阴性，99%的人是真阴性
  - 这种X疾病的正常比例是不大的，1000个人中只有一个人有X病，即0.1%的发病率
- 医生的计算方法：
  - 因为测试的误报率是1%，1000个人将有10个被报为“假阳性”，而根据X病在人口中的比例为1/1000，真阳性只有1个
  - 所以，大约11个测试为阳性的人中只有一个是真阳性（有病）的，因此，王某得病的几率是大约1/11，即0.09(9%)

从一个假想案例说起

张天蓉博主博文中提到的一个概率问题引发了科学网博客上关于贝叶斯定理的大讨论，精彩纷呈，妙趣横生，让观者如痴如醉。

A：普通人群中的王宏感染X病
B：阳性结果

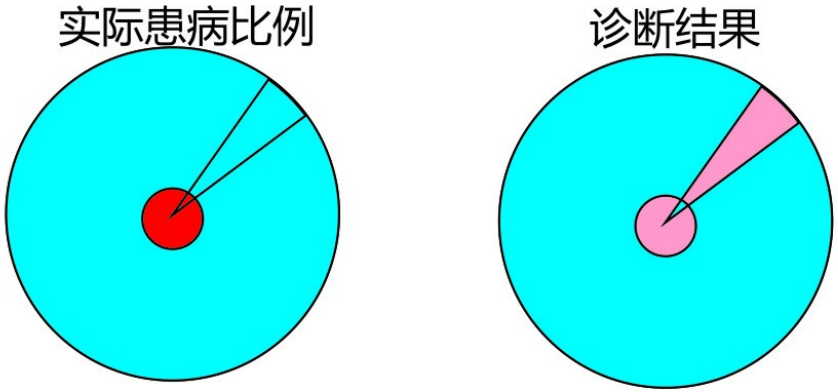
$P(A)$	普通人群中感染X病的概率
$P(B\|A)$	阳性结果的正确率
$P(A\|B)$	有了阳性结果的条件下，王宏感染X病之概率
$P(B)$	结果为阳性的总可能性 = 检查阳性中的真阳性 + 检查阴性中的真阳性

$$P(A|B) = \frac{P(B|A)}{P(B)} P(A) = \frac{99\%}{99\% * (1/1000) + 1\% * (999/1000)} \times (1/1000)$$

$$= \frac{99}{1098} = 9\%$$

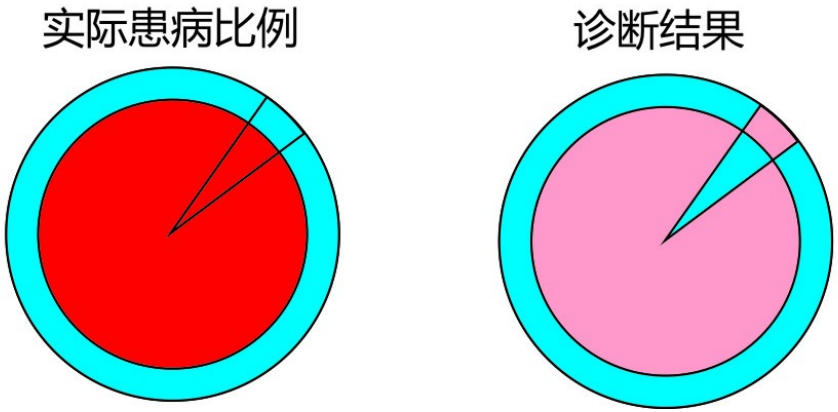
从一个假想案例说起

罕见病症



误诊率1%

流行病症



从一个假想案例说起

我们之前听说，好莱坞影星安吉丽娜朱莉通过《纽约时报》向大家揭露了她惨痛的经历。因携带有BRCA1/2致癌基因，她有87%的机会患上乳腺癌（我们假设5年）。为了防患于未然，她决定切除双乳(确切说是切除双侧乳腺)。美国乳腺癌的发病率为246,680/104,442,302=0.236%（每年每1000成年妇女中有2.36个人患上癌症。我们把预防癌症定位为5年不得癌症。

A:普通妇女5年癌症得病；

B：BRCA1/2致癌基因阳性；

P(A)：普通妇女人群5年患癌概率1.18%；

P(B｜A)：乳腺癌病人中BRCA1/2致癌基因阳性检测率87%；

P(A｜B)：有阳性结果的条件下，安吉丽娜朱莉5年内患癌概率；

P(B)：结果为阳性的总可能性=检查阳性中的真阳性+检查阴性中的真阳性。

通过贝叶斯公式，P(A｜B)= P(B｜A)×P(A)/ P(B)=87%×1.18%/（87%×1.18%+13%×98.2%）=7.4%。

Bayesian Learning

- Bayes rule:

Prior Likelihood

Posterior → $$P(\theta \mid D) = \frac{P(\theta)P(D \mid \theta)}{P(D)}$$ ← Data distribution

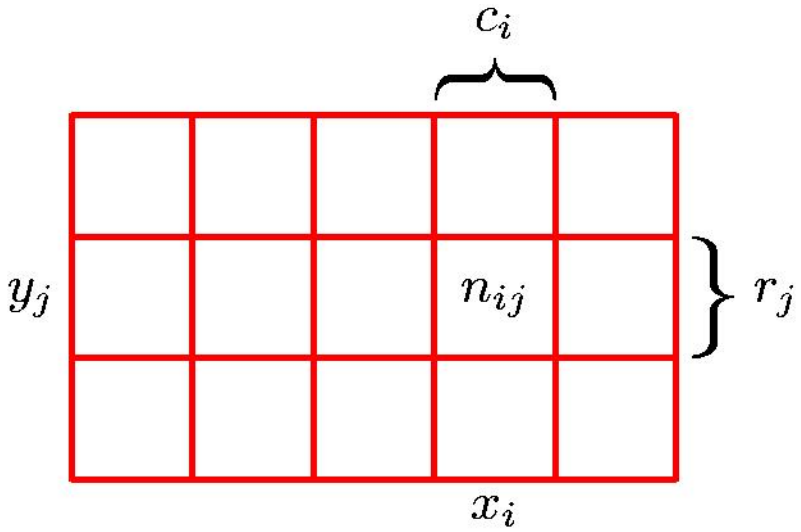
evidence

(Normalization)

- Or equivalently:

$$P(\theta \mid D) \propto P(\theta)P(D \mid \theta)$$

Probability Theory



- Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

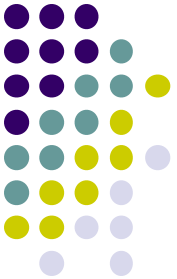
$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

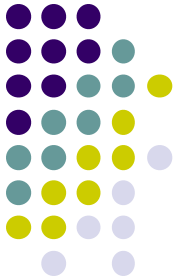
$$= p(Y = y_j | X = x_i)p(X = x_i)$$

Probability concepts



- Random variables: x
- Probability (function): $P(X \leq x)$, $P(x)$
- Density (function): $f(x)$,
- Independency: $P(x, y)=P(x)P(y)$
- Feature quantities:
  - Mean, expectation $E(x) = \int x\, f(x)\, \mathrm{d}x$
  - Covariance
    - $\mathrm{cov}(x,y)=0$, uncorrelatedness / irrelevant (统计无关)
  - Higher order moments

The Rules of Probability



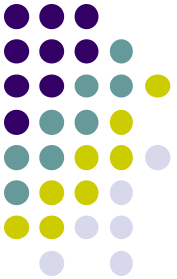
- Sum Rule

$$p(X) = \sum_{Y} p(X, Y)$$

- Product Rule

$$p(X, Y) = p(Y|X)p(X)$$

Bayes' Theorem

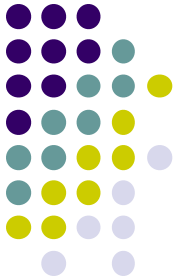


$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

posterior ∝ likelihood × prior

Bayesian Learning in our case



- Likelihood function is simply Binomial:

$$P(D \mid \theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors: （共轭先验）
  - Closed-form representation of posterior
  - For Binomial, conjugate prior is Beta distribution

Beta prior distribution – P(θ)

Gamma function

$$\Gamma(x+1) = x\Gamma(x), \Gamma(1) = 1$$

- Prior: Beta distribution

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\theta \mid \beta_H, \beta_T) = \frac{\Gamma(\beta)}{\Gamma(\beta_H)\Gamma(\beta_T)}\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}$$

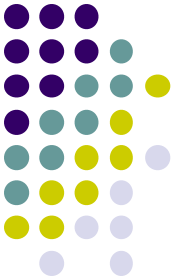
- Likelihood: Binomial distribution

$$P(D \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

- Posterior:

$$\begin{aligned} P(\theta \mid D) &\propto P(\theta)P(D \mid \theta) \\ &\propto \theta^{\alpha_H}(1-\theta)^{\alpha_T}\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1} \\ &\sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T) \end{aligned}$$

Using Bayesian posterior



- Posterior distribution:

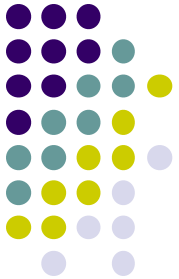
$$P(\theta \mid D) \sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

- Bayesian inference:
  - No longer single parameter:

$$E[f(\theta)] \sim \int_0^1 f(\theta) P(\theta \mid D) d\theta$$

  - Integral, ☹

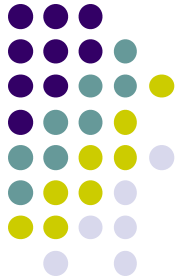
Expectation (数学期望)



- Random variable: θ
- Random function: $f(\theta)$
- Expectation:

$$E[f(\theta)] \sim \int_0^1 f(\theta) P(\theta \mid D) d\theta$$

MAP:
Maximum a posteriori approximation



$$P(\theta \mid D) \sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

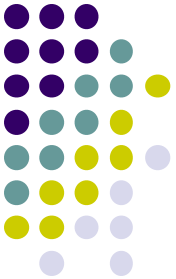
$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid D) d\theta$$

approximation

- MAP: use most likely parameter

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid D) \qquad E[f(\theta)] \approx f(\hat{\theta})$$

MAP for Beta distribution



$$P(\theta \mid D) \sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

- MAP: use most likely parameter

$$\hat{\theta} = \arg\max_{\theta} P(\theta|D) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

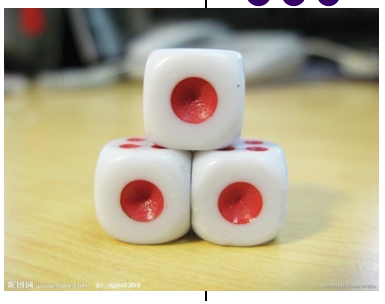
- Beta prior equivalent to extra thumbtack flips
- As $N = \alpha_T + \alpha_H \to \infty$, prior is "forgotten"
- But, for small sample size, prior is important!

More …

- B: Can we handle more complex cases?
- Y: Yes, :-D

- Prior: a mixture of beta distribution
  - $$P(\theta) \sim 0.4Beta(20,1) + 0.4Beta(1,20) + 0.2Beta(2,2)$$

Multinomial distribution



- B: Now if I give you a dice (骰子), then …
- Y: I can solve this problem in a similar way.
- Likelihood:

$$P(X = x^k \mid \boldsymbol{\theta}) = \theta_k,\ k = 1, 2, ..., r,$$

$$\boldsymbol{\theta} = \{\theta_1, ..., \theta_r\},\ \theta_1 + ... + \theta_r = 1$$

$$D = \{X_1 = x_1, ..., X_N = x_N\} => \{N_1, ..., N_r\}$$

$$P(D \mid \boldsymbol{\theta}) = \prod_{i=1}^{r} \theta_i^{N_i}$$

Multinomial distribution

- Conjugate prior (Dirichlet distribution):

$$P(\boldsymbol{\theta}) = \mathrm{Dir}(\ \ |\alpha_1,\ldots,\alpha_r) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{r}\Gamma(\alpha_k)}\prod_{k=1}^{r}\theta_k^{\alpha_k-1}, \quad \alpha = \sum_{k=1}^{r}\alpha_k$$

- Solution:

$$P(X_{N+1} = x^k \mid D) = \int \theta_k \mathrm{Dir}(\boldsymbol{\theta} \mid \alpha_1 + N_1,\ldots,\alpha_r + N_r)d\ \ = \frac{\alpha_k + N_k}{\alpha + N}$$

- Important fact:

$$P(D) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)}\prod_{k=1}^{r}\frac{\Gamma(\alpha_k + N_k)}{\Gamma(\alpha_k)}$$

LDA: Topic Models

Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03

Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng
Computer Science Depa
Stanford University
Stanford, CA 94305, US

Michael I. Jordan
Computer Science Divis
University of California
Berkeley, CA 94720, US

Editor: John Lafferty

[PDF] **Latent dirichlet allocation**

DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org

We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in …

☆ Save 99 Cite Cited by 41445 Related articles All 108 versions »

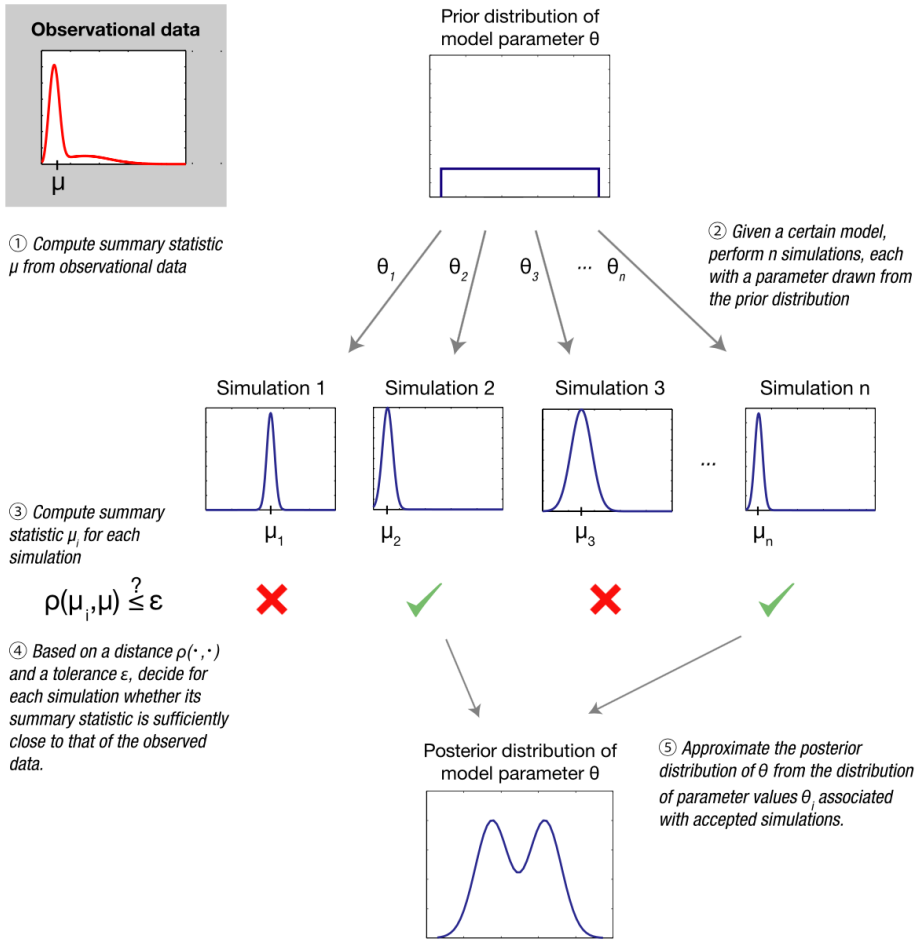
Abstract

We describe *latent Dirichlet allocation* (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

Beyond MAP: Approximate Bayesian Computation

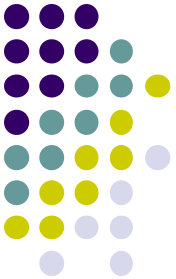
请参考
《近似 Bayes 计算前沿研究进展及应用》

- Approximate Bayesian Computational methods [ARXIV]
- Approximate Bayesian Computation (ABC) in practice [PDF]



Beyond MAP: Approximate Bayesian Computation

请参考
《近似 Bayes 计算前沿研究进展及应用》



算法 1　ABC 算法

算法输入:

(I1) 观测数据 y_n.

(I2) 模型参数 θ 的先验分布 $\pi(\theta)$.

(I3) 数据生成器 $G(z \mid \theta)$.

(I4) 观测数据统计量 $\eta(y_n)$.

(I5) 观测数据统计量的距离度量 $D(\eta(z),\eta(z'))$.

(I6) 距离门限值 $\delta > 0$.

算法流程:

(S1) 从先验分布 $\pi(\theta)$ 中抽取一个候选参数 θ^*.

(S2) 从数据生成器 $G(z \mid \theta^*)$ 中产生一个合成数据 z,并计算其统计量 $\eta(z)$.

(S3) 计算合成数据统计量 $\eta(z)$ 与观测数据统计量 $\eta(y_n)$ 之间的距离 $D(\eta(z),\eta(y_n))$,如果 $D(\eta(z),\eta(y_n)) \leqslant \delta$ 则将 θ^* 接受并保留,否则丢弃.

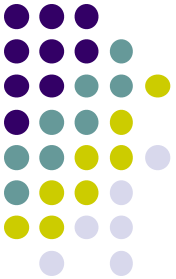
(S4) 重复步骤(S1)~(S3),直到算法的终止条件被满足.

算法输出:

(O1) 在算法运行过程中被保留下来的一组参数样本 $\theta_1^*,\theta_2^*,\cdots,\theta_m^*$.

Beyond MAP: Approximate Bayesian Computation

请参考
《近似 Bayes 计算前沿研究进展及应用》



ABC 算法主要包括以下几个步骤：

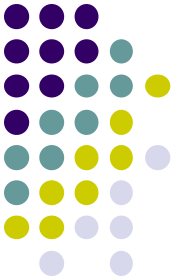
1. **先验分布抽样**：从参数的先验分布 $P(\theta)$ 中抽取一个参数 θ。这就好比根据我们已有的知识或假设，先对参数的可能值做一个初步的猜测。
2. **模拟数据**：使用抽到的参数 θ，根据模型生成一个模拟数据集 D'。也就是说，基于模型和抽到的参数，创造出一组模拟的数据。
3. **计算距离**：计算模拟数据 D' 和实际数据 D 之间的距离，可以用像 $\rho(D, D')$ 这样的度量指标。这一步是为了衡量模拟出来的数据和真实数据有多相似。
4. **接受或拒绝**：如果 $\rho(D, D')$ 小于某个阈值 ϵ，就接受这个参数 θ；否则，就拒绝它。也就是说，如果模拟数据和真实数据足够接近，我们就把 θ 当作可能是真实参数值的一个候选。
5. **重复多次**：重复上述步骤很多次，用接受下来的 θ 值构建出一个近似的后验分布。通过多次重复这个过程，我们可以得到一组和观测数据一致的参数，这组参数就大致代表了参数的后验分布。

选择合适的距离度量指标和阈值 ϵ 非常关键。距离度量指标决定了我们如何衡量模拟数据和真实数据的相似度，阈值 ϵ 则决定了我们接受或拒绝参数的标准。如果 ϵ 太小，可能会导致很少有参数被接受，算法效率低下；如果 ϵ 太大，又可能会导致后验分布的近似效果不好。

通过Deepseek官网应用生成
（打开了深度思考与网络搜索功能，仅供直观理解参考）

Beyond MAP: Approximate Bayesian Computation

请参考
《近似 Bayes 计算前沿研究进展及应用》

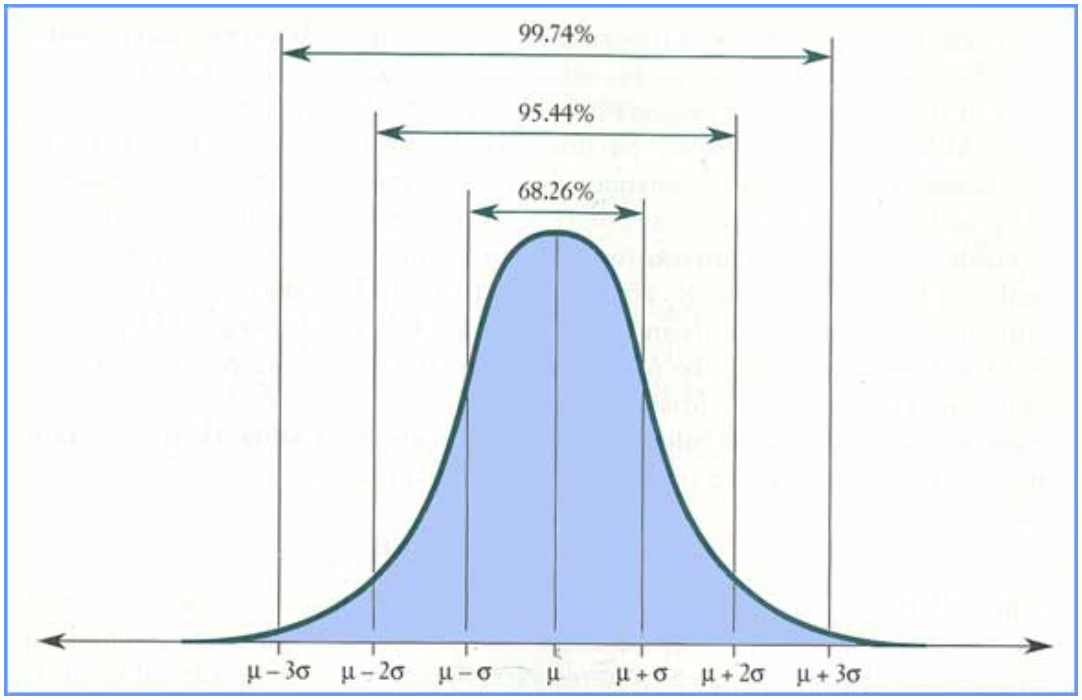


假设我们要估计一个正态分布的均值。我们有一组数据 $D = \{x_1, x_2, \ldots, x_n\}$，它来自一个均值为 μ、方差为 σ^2（已知）的正态分布。

- **MAP 方法**：数据的似然函数为 $L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$。我们可以最大化这个似然函数来得到 μ 的 MAP 估计值。MAP 估计值为 $\hat{\mu}_{MAP} = \bar{x}$，其中 $\bar{x}$ 是样本均值。
- **ABC 算法**：我们可以使用 ABC 算法来近似 μ 的后验分布。首先，从 μ 的先验分布 $P(\mu)$ 中抽取一个参数 μ。然后，从均值为 μ、方差为 σ^2 的正态分布中模拟出一个数据集 D'。计算 D' 和 D 之间的距离，比如欧几里得距离 $\rho(D, D') = \sqrt{\sum_{i=1}^{n}(x'_i - x_i)^2}$。如果 $\rho(D, D')$ 小于阈值 ϵ，就接受 μ；否则，就拒绝它。重复这个过程很多次，得到一组被接受的 μ 值，这组值就大致代表了 μ 的后验分布。

在这个例子中，ABC 算法可以在不计算似然函数的情况下，提供 μ 的近似后验分布，而 MAP 方法如果似然函数能够正确指定，可以提供准确的 μ 估计值。然而，在更复杂的模型中，似然函数可能很难计算，这时 ABC 算法就更有用了。

通过Deepseek官网应用生成
（打开了深度思考与网络搜索功能，仅供直观理解参考）



在数据处理中，高斯分布无处不在

Gaussian distribution

Gaussian distribution

均值

Continuous random variable:

mean

$$P(x \mid \mu, \delta) \sim \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\delta^2}}$$

variance

standard deviation

Normalization

方差　　标准差

Consider the difference between continuous and discrete variables?

MLE for Gaussian

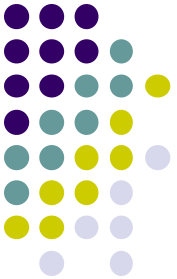
- Prob. of i.i.d. samples $D = \{x_1, x_2, \boxed{?}, x_N\}$

likelihood $$P(D \mid \mu, \sigma) = \left(\frac{1}{\sigma \sqrt{2\pi}} \right)^N \prod_{i=1}^{N} e^{\frac{-(x_i - \mu)^2}{2\sigma^2}}$$

- The magic of log (to log-likelihood)

$$\begin{aligned} \ln P(D \mid \mu, \sigma) &= \ln \left(\frac{1}{\sigma \sqrt{2\pi}} \right)^N \prod_{i=1}^{N} e^{\frac{-(x_i - \mu)^2}{2\sigma^2}} \\ &= -N \ln(\sigma \sqrt{2\pi}) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \end{aligned}$$

MLE for mean of a Gaussian



$$\frac{\partial}{\partial \mu} \ln P(D \mid \mu, \sigma) = \frac{\partial}{\partial \mu} \ln \left(\frac{1}{\sigma \sqrt{2\pi}} \right)^N \prod_{i=1}^{N} e^{\frac{-(x_i - \mu)^2}{2\sigma^2}}$$

$$= \frac{\partial}{\partial \mu} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

$$= \sum_{i=1}^{N} \frac{(x_i - \mu)}{\sigma^2} = 0$$

$$\mu = \frac{1}{N} \sum_i x_i$$

MLE for variance of a Gaussian

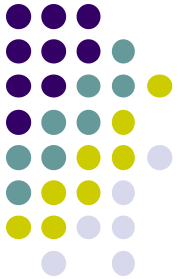
$$\frac{\partial}{\partial\sigma}\ln P(D\mid\mu,\sigma) = \frac{\partial}{\partial\sigma}\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{N}\prod_{i=1}^{N}e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$= \frac{\partial}{\partial\sigma}[-N\ln\sigma\sqrt{2\pi}]-\sum_{i=1}^{N}\frac{\partial}{\partial\sigma}[\frac{(x_i-\mu)^2}{2\sigma^2}]$$

$$= -\frac{N}{\sigma}+\sum_{i=1}^{N}\frac{(x-\mu)^2}{\sigma^3} = 0$$

$$\sigma^2 = \frac{1}{N}\sum_{i}(x_i-\mu)^2$$

Gaussian parameters learning

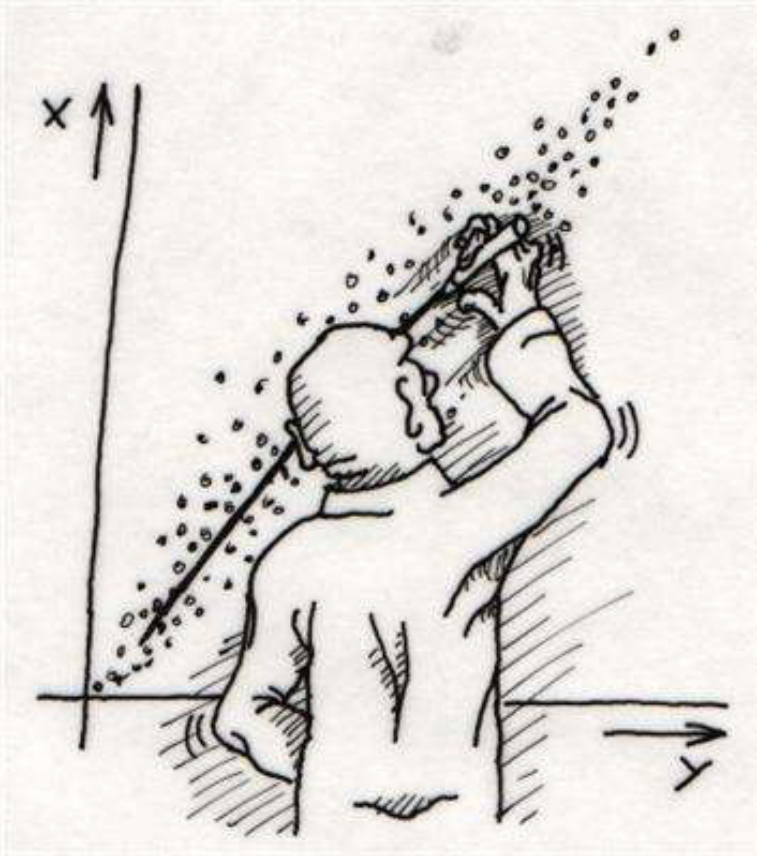


- MLE

$$\hat{\mu} = \frac{1}{N}\sum_i x_i$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i - \mu)^2$$

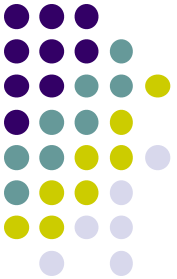
- Bayesian learning: prior?
- Conjugate priors:
  - Mean: Gaussian priors
  - Variance: Wishart Distribution



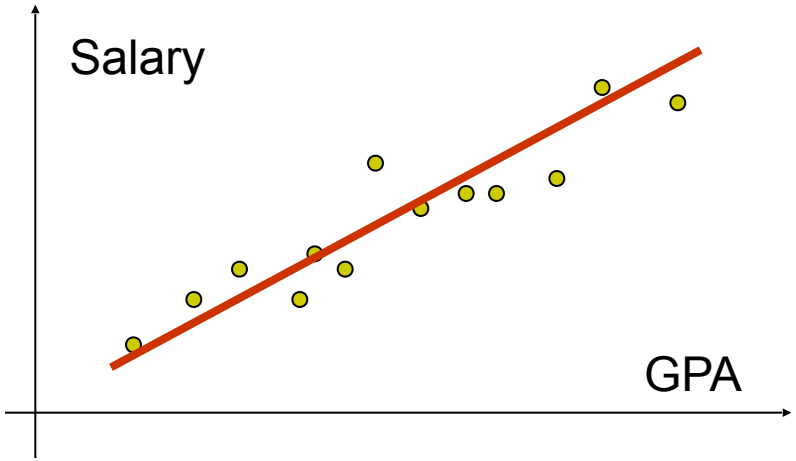
回归的思想，对于理解数据本身的特点与规律，是一种利器

Regression problems

Prediction of continuous variable



- B: Wait, that's not what I meant!
- Y: Chill out, dude.
- B: I want to predict a continuous variable for continuous inputs: I want to predict salaries from GPA.
- Y: I can regress that…

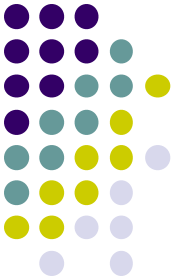


The regression problem

- **Instances:** $< \mathbf{x}_i, t_i >$
- **Learn:** mapping from $\mathbf{x}$ to $t(\mathbf{x})$.
- **Hypothesis space:** $t(\mathbf{x}) \approx \hat{f}(x) = \sum_{i=1}^{k} w_i h_i$
  - Given, basis functions $H = \{h_1, \ldots, h_k\}$
  - Find coefficients $\mathbf{w} = \{w_1, \ldots, w_k\}$
- **Problem formulation:**

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j [t(\mathbf{x}_j) - \sum_{i=1}^{k} w_i h_i(x)]^2$$

But, why sum squared error?

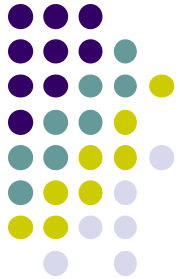


- Model:

$$P(t \mid \mathbf{x}, \mathbf{w}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-[t - \sum_i w_i h_i(x)]^2}{2\sigma^2}}$$

- Learn **w** using MLE

Maximizing log-likelihood

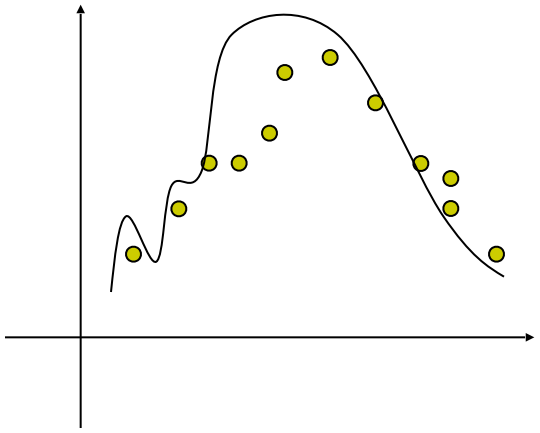


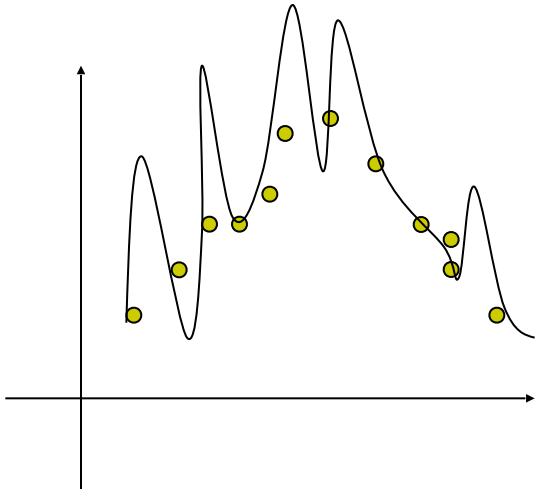
$$\ln P(D \mid \mathbf{w}, \sigma) = \ln \prod_j \left(\frac{1}{\sigma \sqrt{2\pi}} e^{\frac{-[t_j - \sum_i w_i h_i(x_j)]^2}{2\sigma^2}} \right)$$

$$\Rightarrow \min \sum_j \frac{-[t_j - \sum_i w_i h_i(x_j)]^2}{2\sigma^2}$$

Bias-Variance Tradeoff

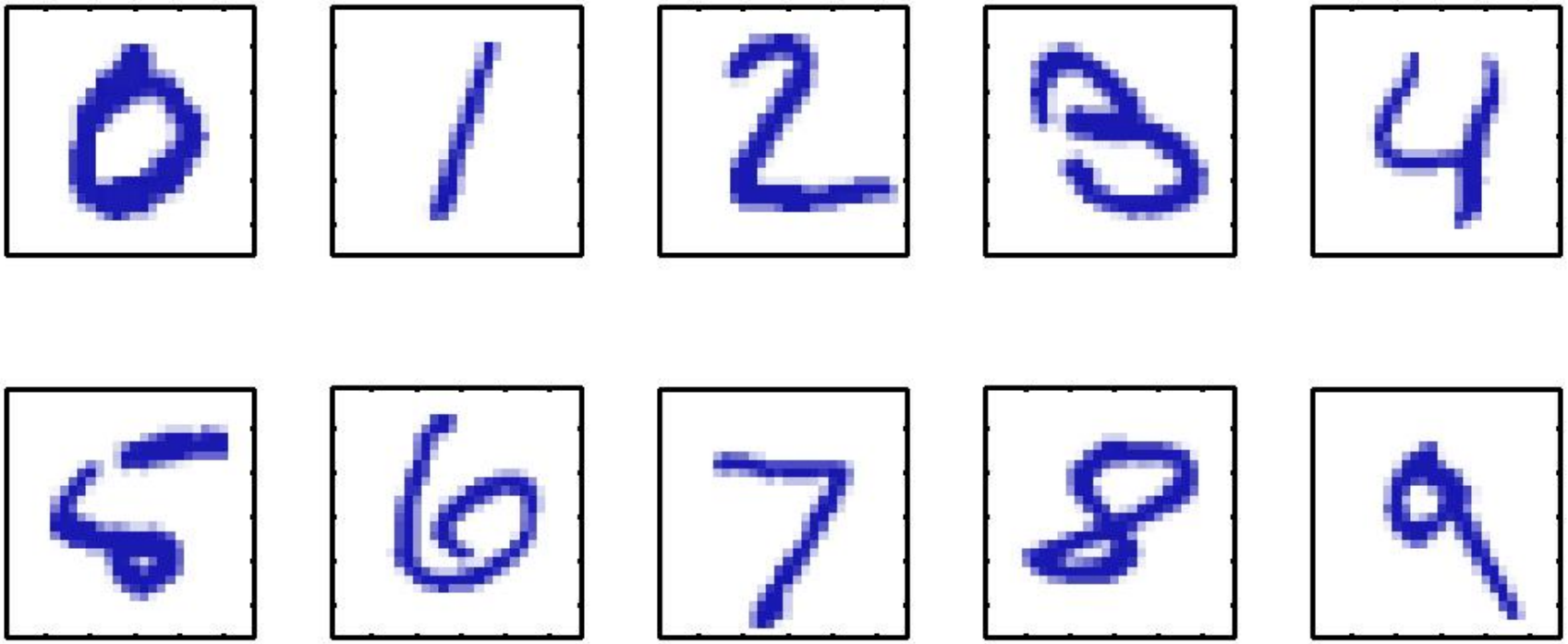
- Choice of hypothesis basis introduce learning bias:
  - More complex basis:
    - Less bias
    - More variance (over-fitting)



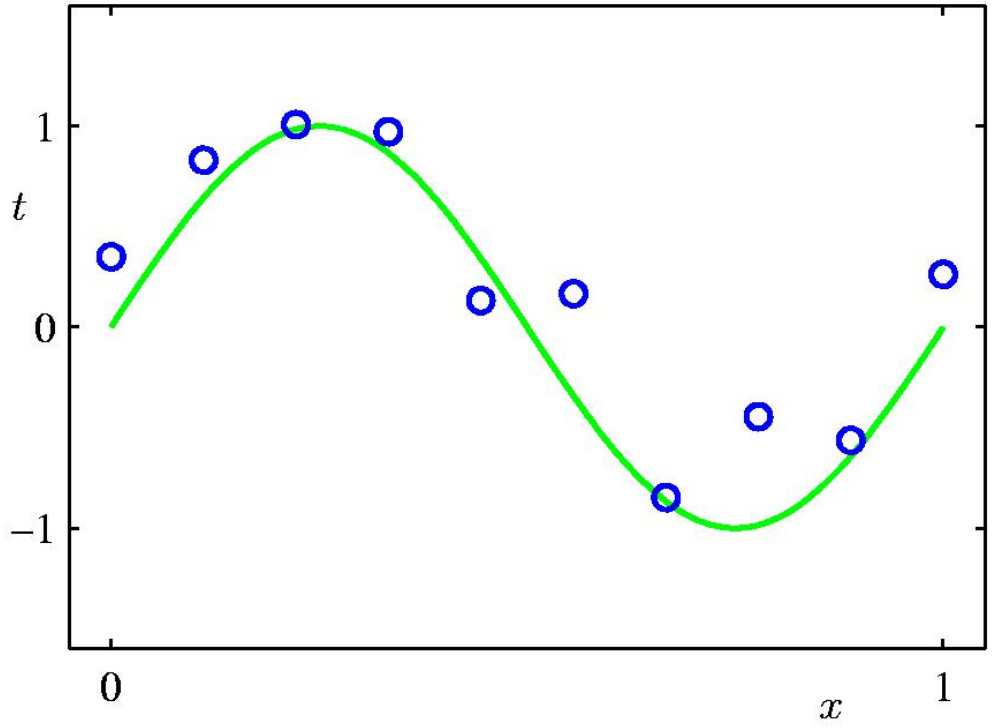


Example

Handwritten Digit Recognition

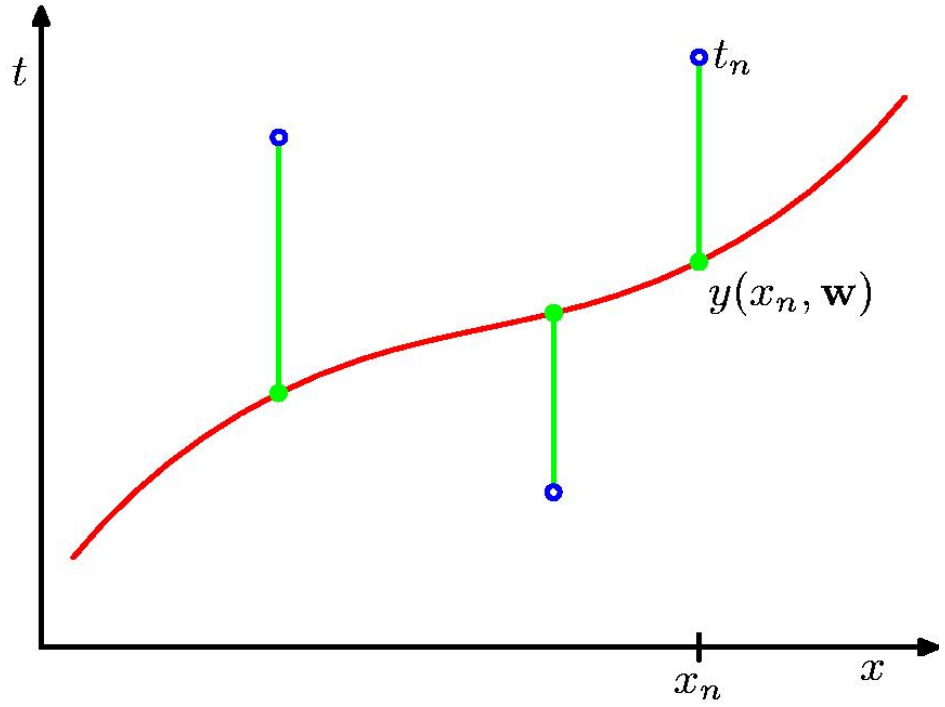


Polynomial Curve Fitting



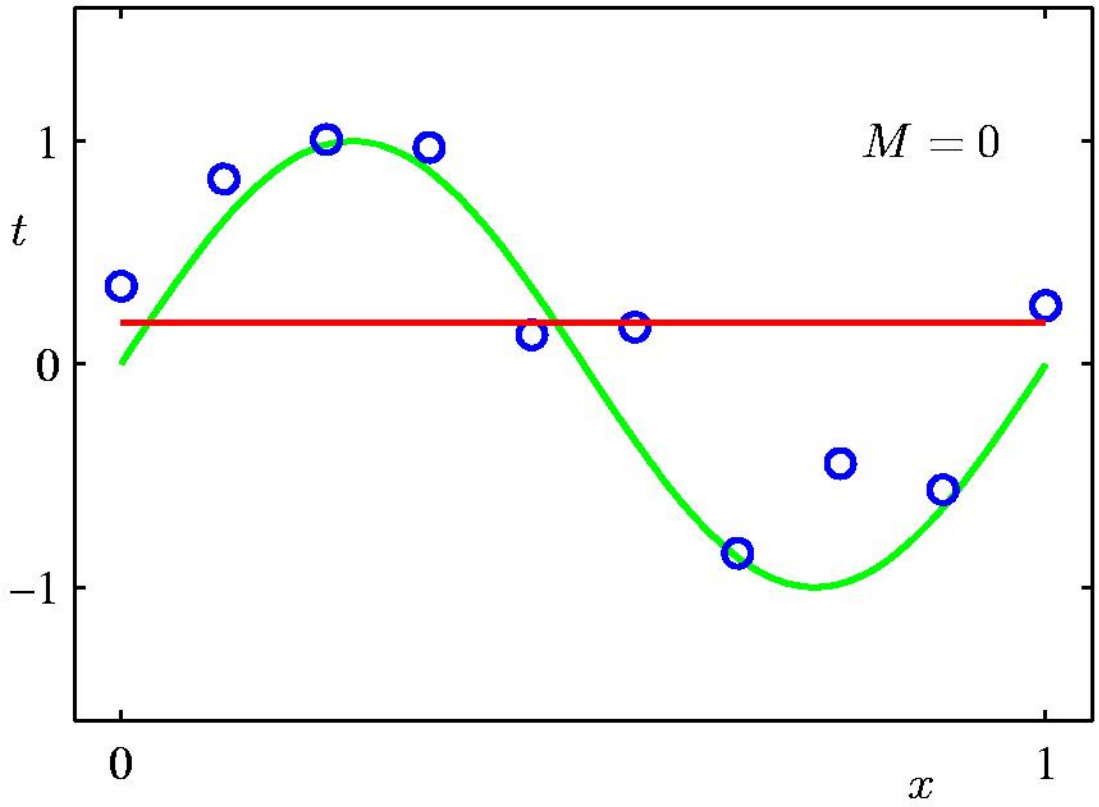
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function

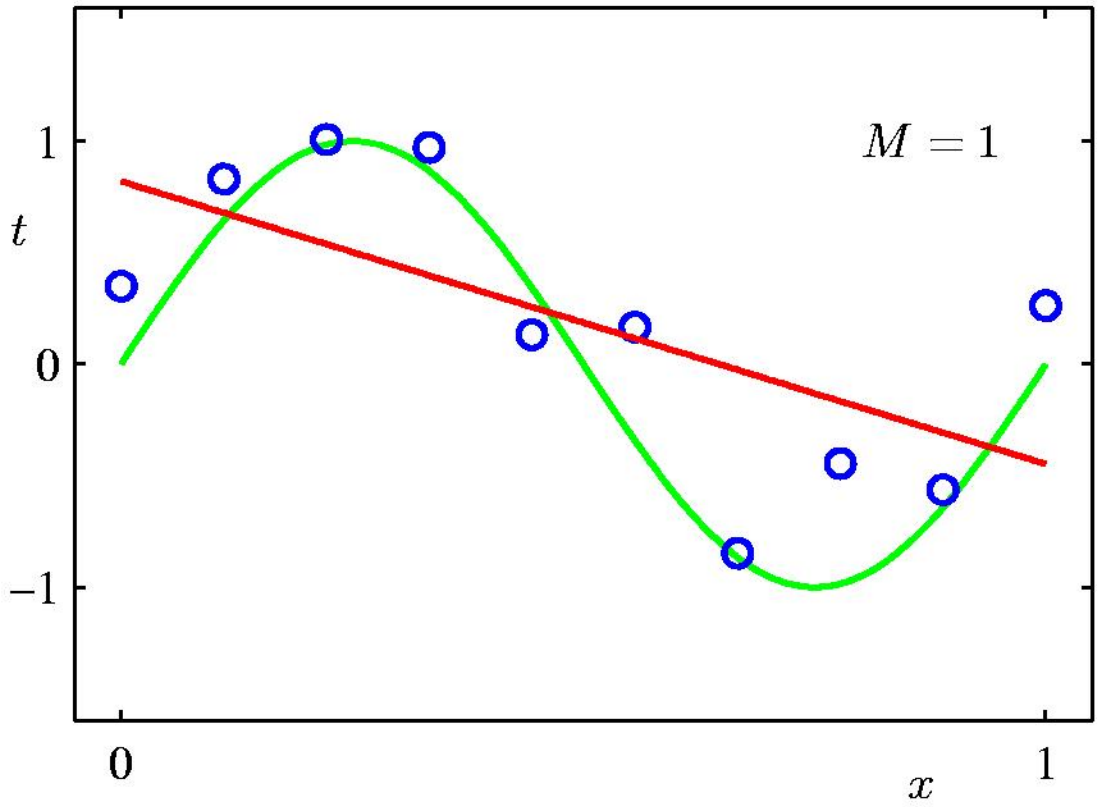


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

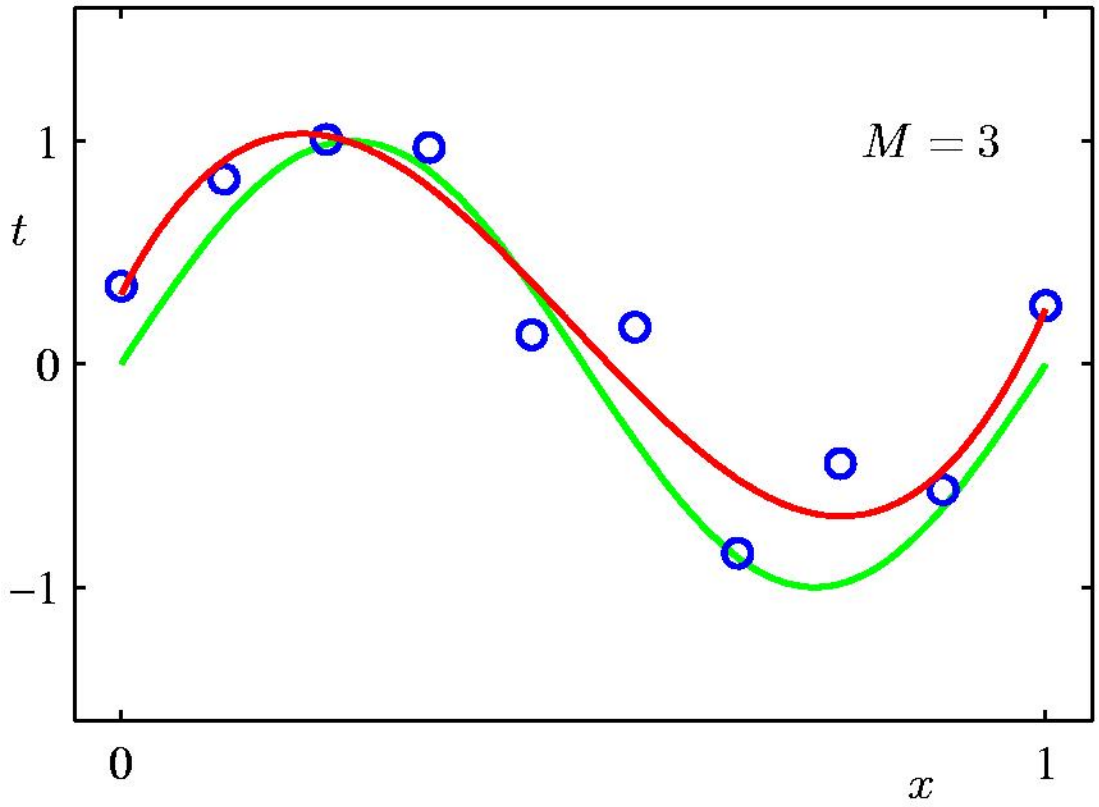
0th Order Polynomial



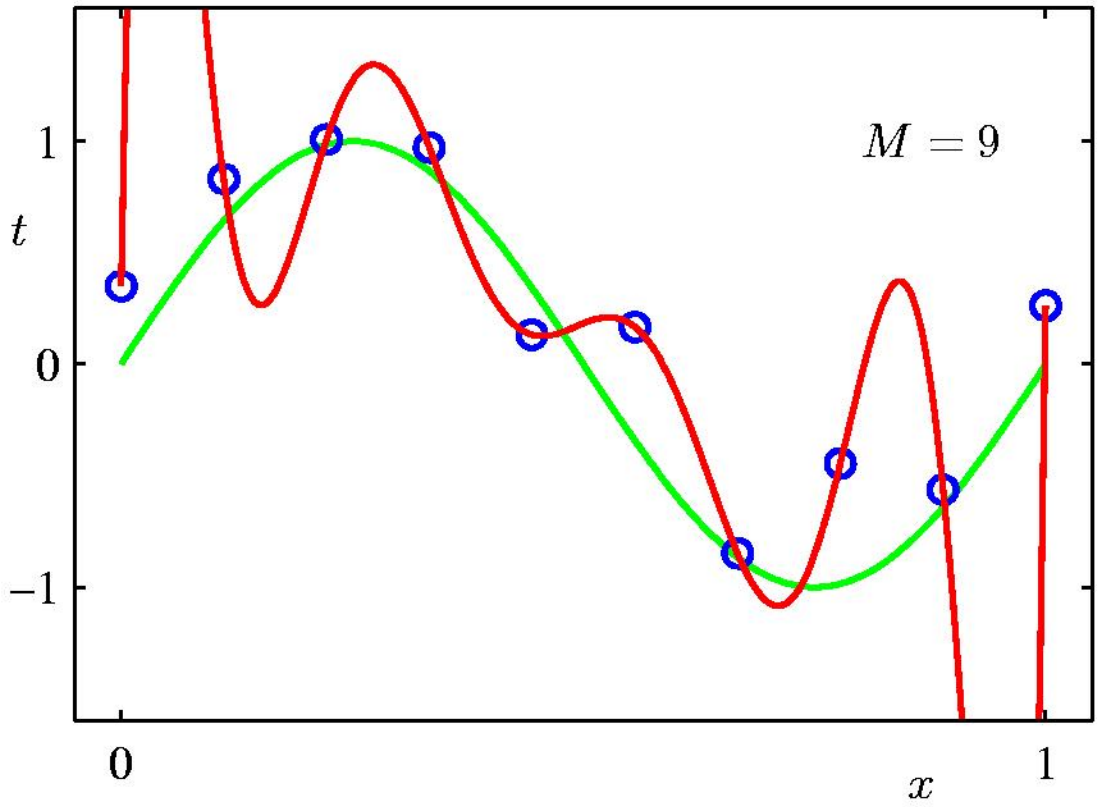
1st Order Polynomial



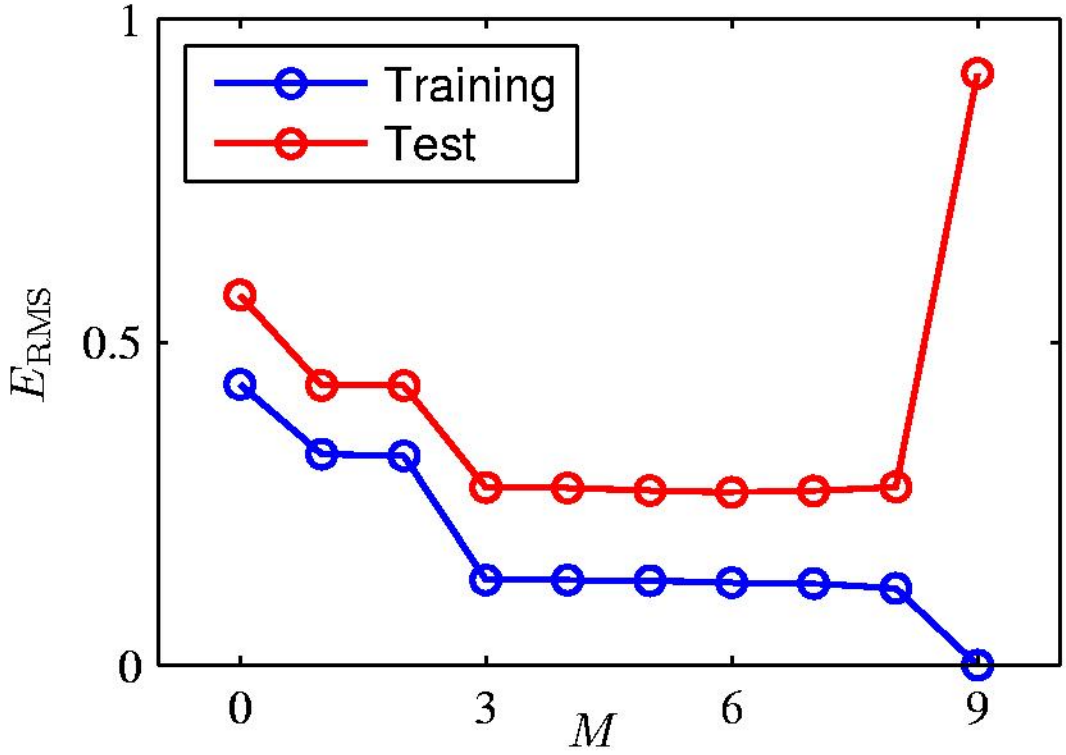
3rd Order Polynomial



9th Order Polynomial

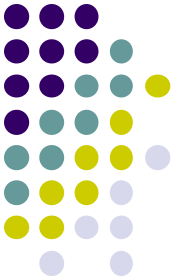


Over-fitting



Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^{\star})/N}$

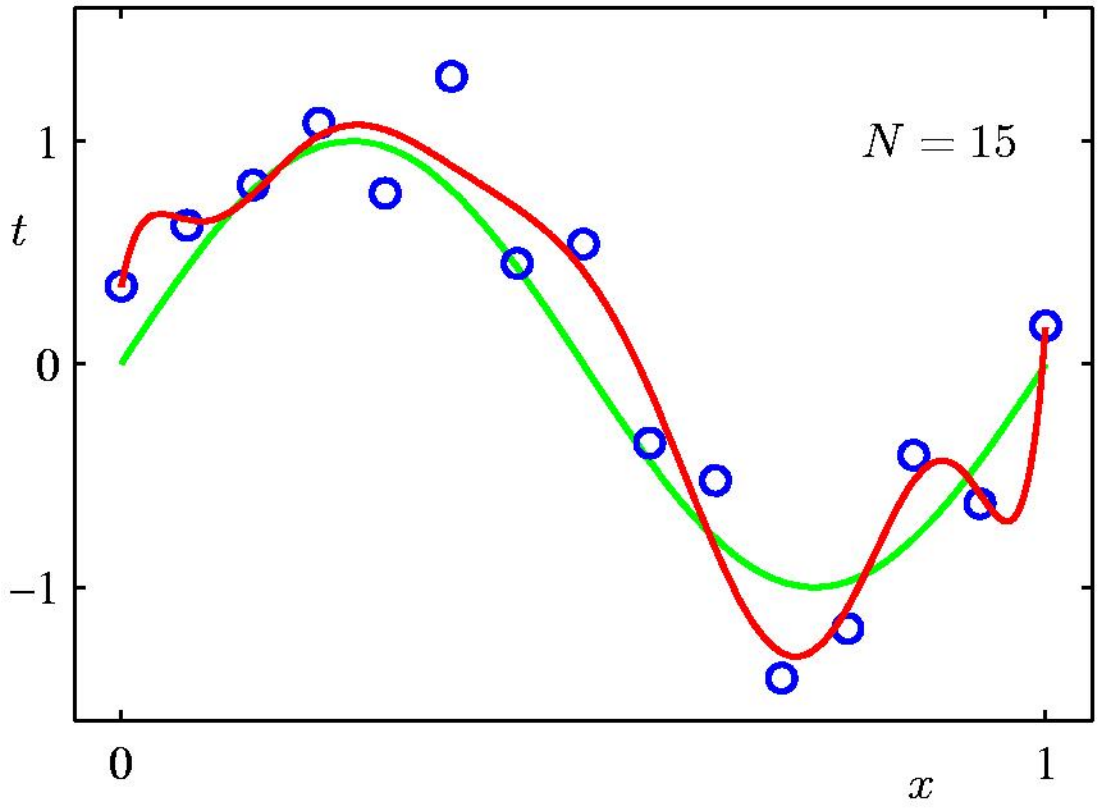
Polynomial Coefficients



	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

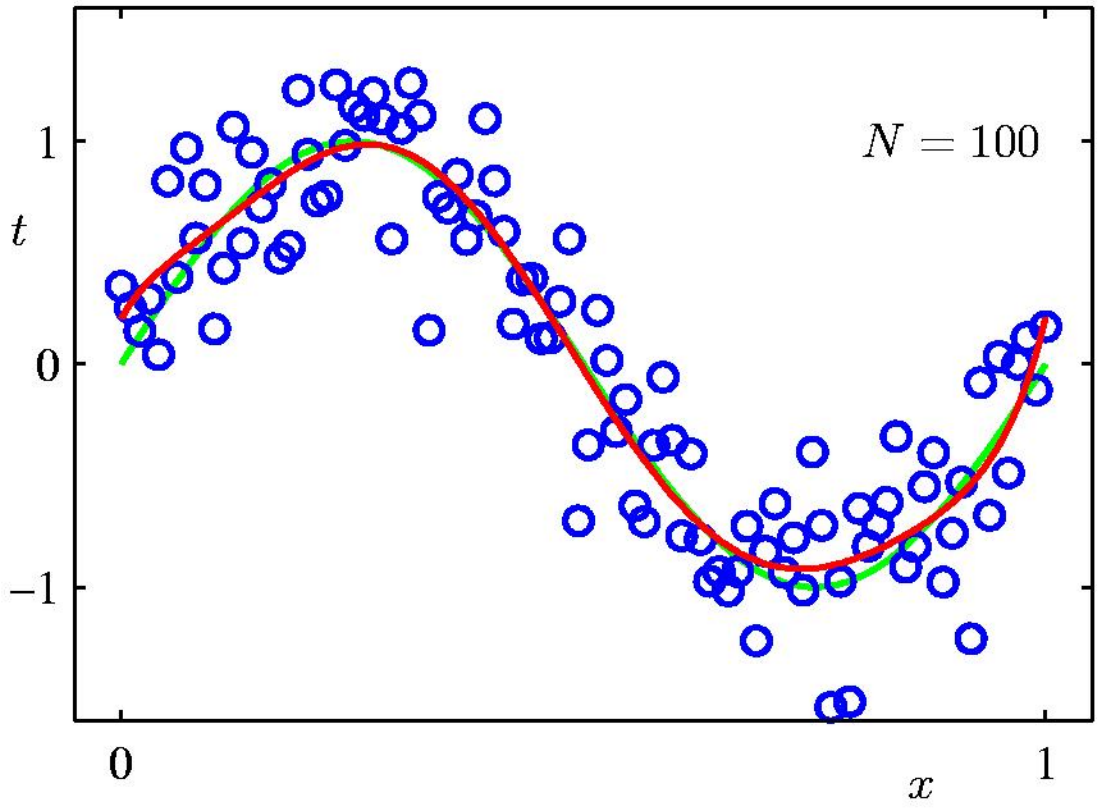
Data Set Size: $N = 15$

9th Order Polynomial

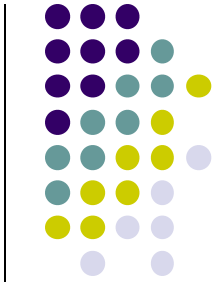


Data Set Size: $N = 100$

9th Order Polynomial



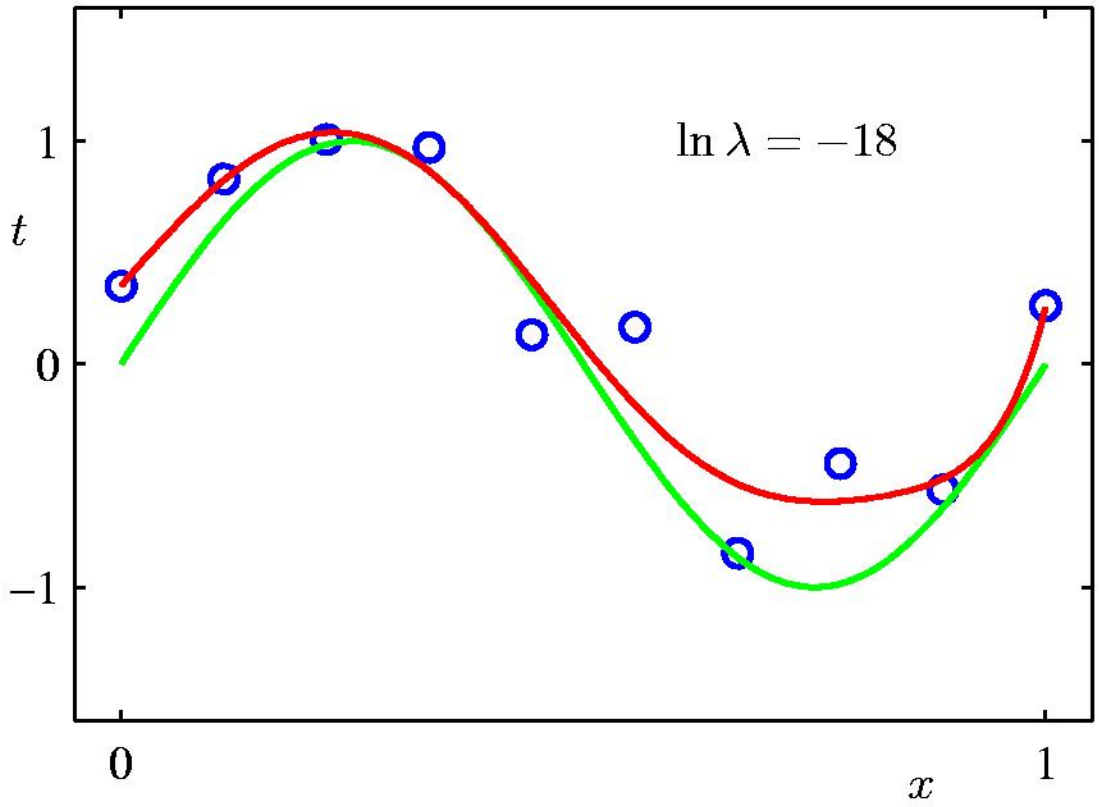
Regularization



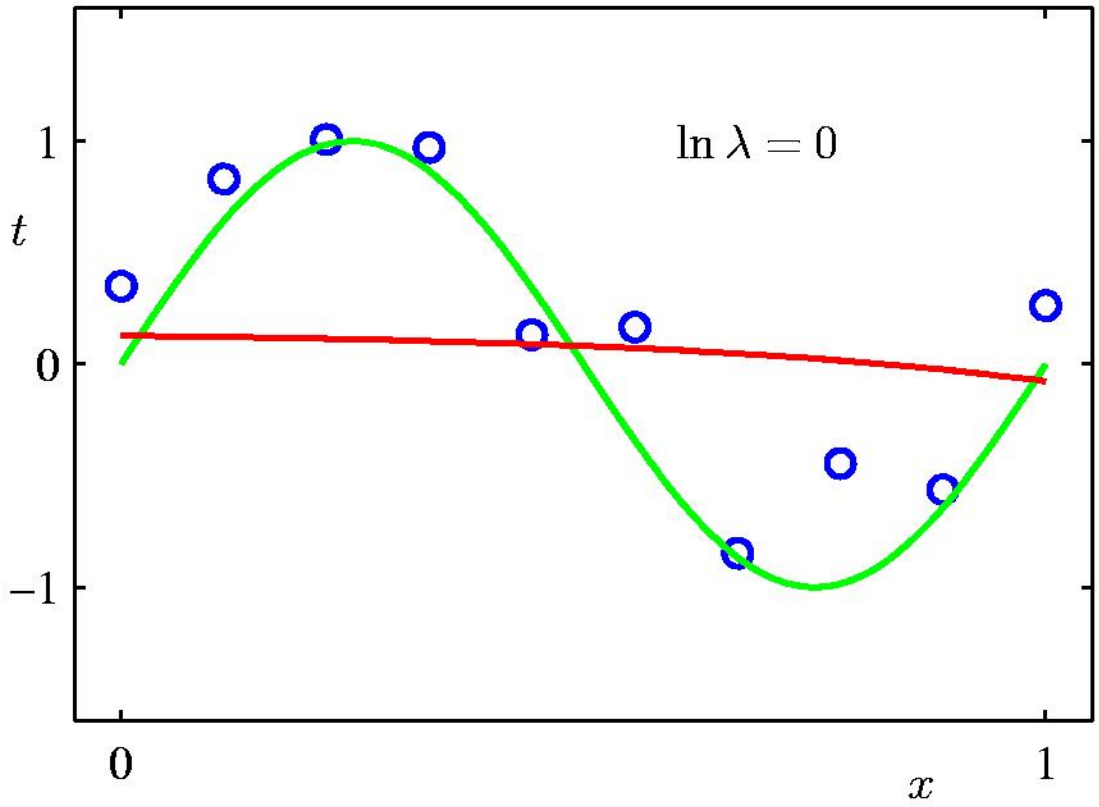
- Penalize large coefficient values

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

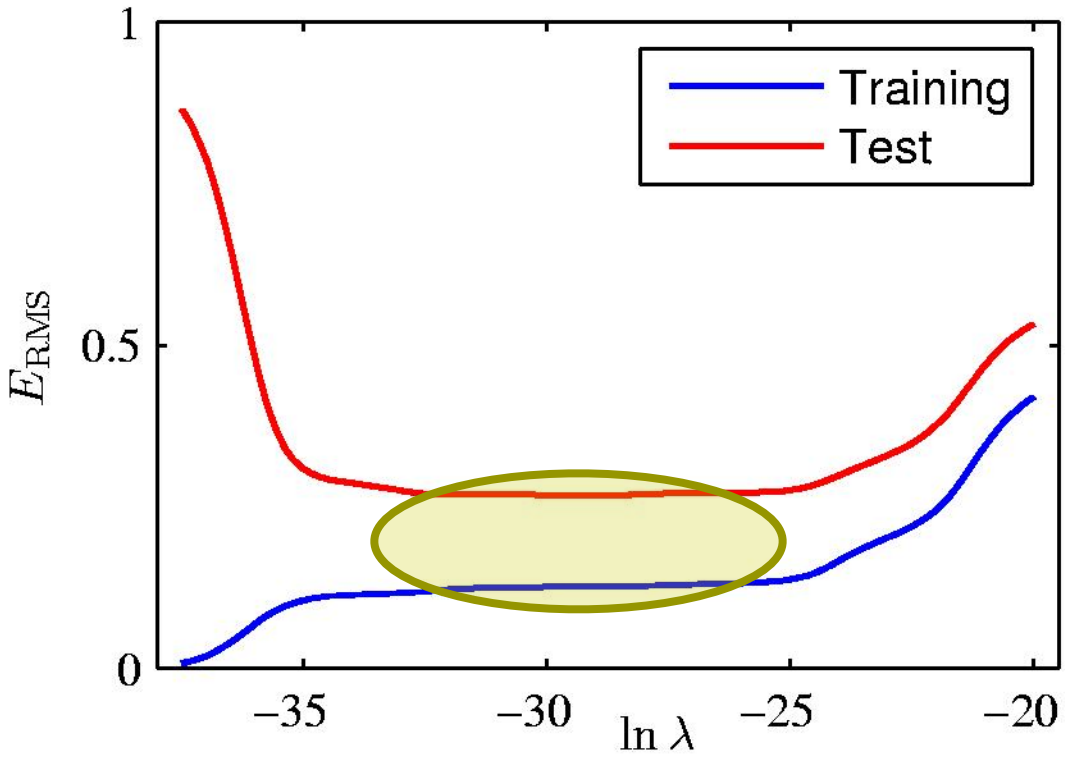
Regularization: $\ln \lambda = -18$



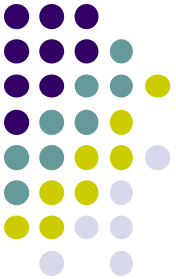
Regularization: $\ln \lambda = 0$



Regularization: E_{RMS} vs. $\ln \lambda$



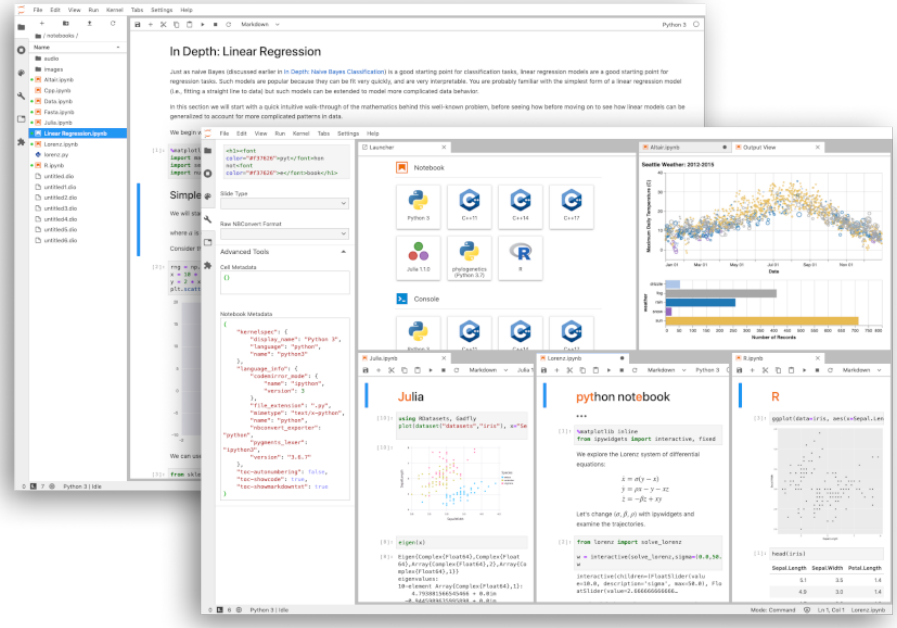
What you need to know



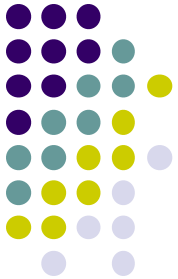
- Point estimation:
  - Maximal Likelihood Estimation
  - Bayesian learning
  - Maximal a Posterior
- Gaussian estimation
- Regression
  - Basis function = features
  - Optimizing sum squared error
  - Relationship between regression and Gaussians
- Bias-Variance trade-off

Homework-01

- Python programming
  - Basic: 1-D regression (polynomial basis)
  - Advance: 1-D regression (Bernstein basis/Bezier)
  - Learn Jupyter Notebook (https://jupyter.org)



Question for Review



- Finish the “Gaussian parameters learning”
  - Please use google, ^_*

微博：@浙大张宏鑫
邮件：zhx@cad.zju.edu.cn
主页：http://person.zju.edu.cn/zhx
手机：13958011790
微信：timothykull

谢谢

Thank You