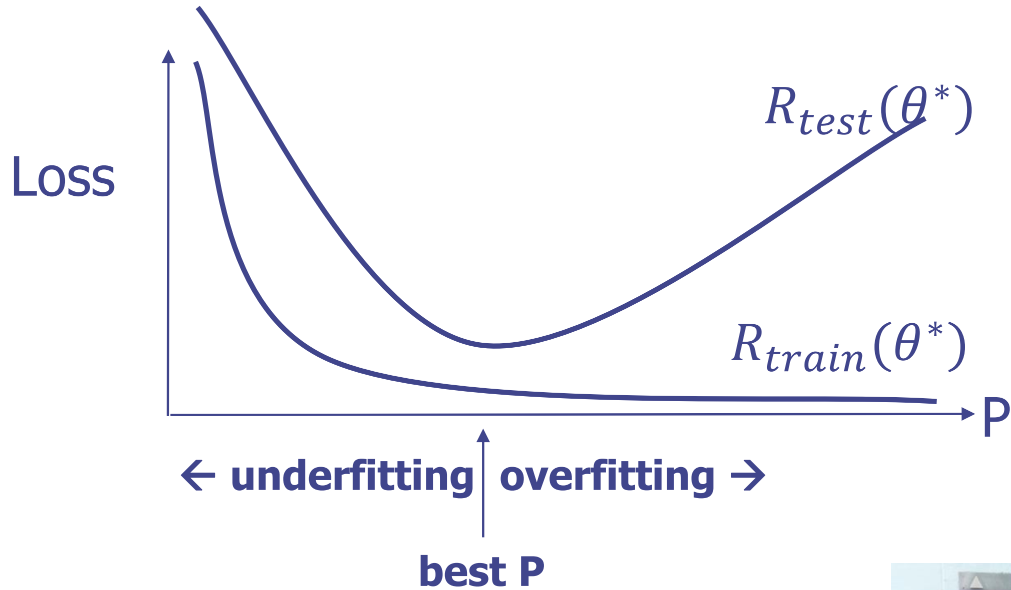
Machine Learning

4771

Instructor: Itsik Pe’er

Reminder: Cross Validation



General Additive Models

Class 5: How to stop Max Likelihood from Overfitting ?

- Estimating parameters of distributions
- Evidence vs. prior assumptions
- Regularizing regression

Example: Mean of Gaussian

- Can we recover most likely μ for height?

$x \sim Normal(\mu, \sigma^2)$

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

$$\log P(x_1, x_2, \ldots x_N | \mu, \sigma^2) = \sum_{n=1}^{N} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

Example: Mean of Gaussian

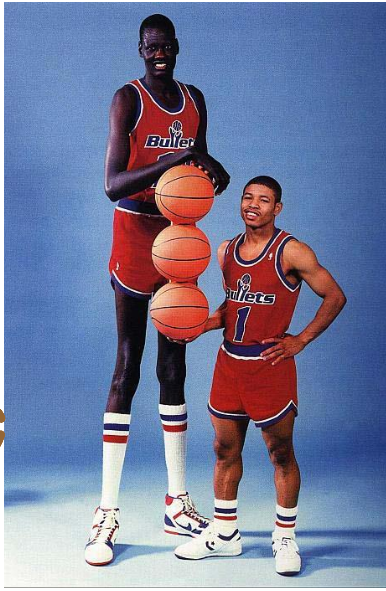
- Can we recover most likely μ for height?

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

$$\log p(x_1, \ldots, x_N|\mu,\sigma^2) =$$

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

$$\frac{d}{d\mu}\log p(x|\mu^*,\sigma^2) = + \frac{\sum \cancel{2}(x_i-\mu^*)}{\cancel{2}\sigma^2} = 0$$



Example: Mean of Gaussian

- Can we recover most likely μ for height?

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

$$\log p(x_1, \ldots, x_N|\mu,\sigma^2) =$$
$$= -\frac{N}{2}\log 2\pi\sigma^2 - \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{d}{d\mu}\log p(\boldsymbol{X}|\mu^*,\sigma^2) = \frac{\sum_{i=1}^{N}(x_i-\mu^*)}{\sigma^2} = 0$$

$$\mu^* = \frac{\sum_{i=1}^{N} x_i}{N}$$

Example: Success rate

- Can we recover ML α for drawing a card?

$$x \sim Bernoulli(\alpha)$$

$$p(x|\alpha) = \alpha^x (1-\alpha)^{1-x}$$

$N_1 = \sum X_i$

$\log P(X|\alpha) = \sum_i \log p(x_i | \alpha) = \sum_{i | x_i = 0} \log(1-\alpha) + \sum_{i | x_i = 1} \log \alpha$

$(N - N_1)\log(1-\alpha) + N_1 \log \alpha$

$\frac{\partial}{\partial \alpha} \log P(X|\alpha^*) = \frac{N_1}{\alpha^*} - \frac{N - N_1}{1-\alpha^*} = 0$

Example: Success rate

- Can we recover ML α for drawing a card?

$$x \sim Bernoulli(\alpha)$$

$$p(x|\alpha) = \alpha^x (1-\alpha)^{1-x}$$

$$N_1 = \sum_i x_i$$

$$\log p(x_1, \ldots, x_N|\alpha) = N_1 \log \alpha - (N - N_1) \log(1-\alpha)$$

Example: Success rate

- Can we recover ML α for drawing a card?

$$x \sim Bernoulli(\alpha)$$

$$p(x|\alpha) = \alpha^x (1-\alpha)^{1-x}$$

$$N_1 = \sum_i x_i$$

$$\log p(x_1, \dots, x_N|\alpha) = N_1 \log \alpha - (N - N_1)\log(1-\alpha)$$

$$\frac{d}{d\alpha} \log p(\boldsymbol{X}|\alpha^*) = \frac{N_1}{\alpha^*} - \frac{N-N_1}{1-\alpha^*} = 0$$

$$\alpha^* = \frac{N_1}{N}$$

Best Guess

$E(\alpha)$

- Given evidence X, what's best guess α?

Best Guess

- Prior assumption about $\alpha : p(\alpha)$
- What's best guess α? $E[\alpha]$
- Given evidence X, what's best guess α?

Bayesian Inference

- Prior assumption about $\alpha : p(\alpha)$

$$E[\alpha]$$

- Given evidence X, what's best guess α?

- Bayesian answer: optimize $E[\alpha|X]$ w.r.t. posterior $p(\alpha|X) = \dfrac{p(\alpha)p(X|\alpha)}{p(X)}$

likelihood

- Optimal if we have true probability

Bayesian Inference

- Prior assumption about $\alpha : p(\alpha)$

  $E[\alpha]$

- Given evidence $\boldsymbol{X}$, what's best guess α?

- Bayesian answer: optimize $E[\alpha|\boldsymbol{X}]$ w.r.t. posterior

prior

$$p(\alpha|\boldsymbol{X}) = \frac{p(\alpha)p(\boldsymbol{X}|\alpha)}{p(\boldsymbol{X})}$$

likelihood

Constant w.r.t. α

Bayesian Inference

- Prior assumption about $\alpha : p(\alpha)$
- Given evidence X, what is the Expected A-Posteriori (EAP) $E_\alpha\left[\frac{p(\alpha)p(\boldsymbol{X}|\alpha)}{p(\boldsymbol{X})}\right]$

Bayesian Inference

- Prior assumption about $\alpha : p(\alpha)$
- Given evidence X, what is the Expected A-Posteriori (EAP) $E_\alpha\left[\frac{p(\alpha)p(\boldsymbol{X}|\alpha)}{p(\boldsymbol{X})}\right]$
- Another approach: Maximum A-Posteriori (MAP) $argmax_\alpha[p(\alpha)p(\boldsymbol{X}|\alpha)] =$
  $= argmax_\alpha[\log p(\alpha) + \log p(\boldsymbol{X}|\alpha)]$

Bayesian Inference

- Prior assumption about $\alpha : p(\alpha)$

$\alpha \sim \text{Uniform}(0,1)$ $\quad p(X|\alpha) = \alpha^{x}(1-\alpha)^{1-x}$

- Given evidence X, what is the Expected A-Posteriori (EAP) $E_\alpha\left[\frac{p(\alpha)p(\boldsymbol{X}|\alpha)}{p(\boldsymbol{X})}\right] =$

$$= \frac{\int_0^1 \alpha \cdot 1 \cdot \alpha^{N_1}(1-\alpha)^{N-N_1}\, d\alpha}{\int_0^1 \alpha^{N_1}(1-\alpha)^{N-N_1}\, d\alpha}$$

Bayesian Inference

- Prior assumption about $\alpha : p(\alpha)$

$$\alpha \sim Uniform(0,1) \;;\; x \sim Bernoulli(\alpha)$$

- Given evidence $\boldsymbol{X}$, what is the Expected A-Posteriori (EAP) $E_\alpha\left[\frac{p(\alpha)p(\boldsymbol{X}|\alpha)}{p(\boldsymbol{X})}\right] =$

$$= \frac{1}{p(\boldsymbol{X})}\int_{\alpha=0}^{1} \alpha\, p(\alpha)p(\boldsymbol{X}|\alpha)d\alpha =$$

$$= \frac{\int_{\alpha=0}^{1} \alpha \cdot 1 \cdot \alpha^{N_1}(1-\alpha)^{N-N_1}d\alpha}{\int_{\alpha=0}^{1} \alpha^{N_1}(1-\alpha)^{N-N_1}d\alpha} = \frac{c(N_1+1, N-N_1)}{c(N_1, N-N_1)}$$

$$c(m,k) = \int_{\alpha=0}^{1} \alpha^m (1-\alpha)^k d\alpha$$

$k=0$:

$$c(m,0) = \frac{\alpha^{m+1}}{m+1}\Bigg|_0^1 = \frac{1}{m+1}$$

$k>0,\ m>0$

$$0 = \alpha^m (1-\alpha)^k \Big|_0^1 = \cancel{m \int_0^1 \alpha^{m-1} (1-\alpha)^k d\alpha} - k \cancel{\int_0^1 \alpha^m (1-\alpha)^{k-1} d\alpha}$$

$c(m-1,k)$ $\quad c(m,k-1)$

$$c(m,k) = \frac{k}{m+1} c(m+1,k-1)$$

$$c(m,k) = \int_{\alpha=0}^{1} \alpha^m (1-\alpha)^k \, d\alpha$$

$$k = 0 : c(m,k) = \int_{\alpha=0}^{1} \alpha^m \, d\alpha = \frac{1}{m+1}$$

$k, m > 0 :$

$$0 = \alpha^m (1-\alpha)^k \Big|_0^1 = mc(m-1,k) - kc(m,k-1)$$

$$c(m,k) = \frac{k}{m+1} c(m+1,k-1) = \cdots =$$

$$= \frac{k!}{\frac{(m+k)!}{m!}} c(m+k,0) = \frac{m!\,k!}{(m+k)!} \int_{\alpha=0}^{1} \alpha^{m+k} \, d\alpha$$

$$= \frac{m!\,k!}{(m+k+1)!}$$

Bayesian Inference

- Prior assumption about $\alpha : p(\alpha)$

  $\alpha \sim Uniform(0,1)$; $x \sim Bernoulli(\alpha)$

- Given evidence X, what is the Expected A-Posteriori (EAP) $E_\alpha\left[\frac{p(\alpha)p(\boldsymbol{X}|\alpha)}{p(\boldsymbol{X})}\right] =$

$$= \frac{c(N_1+1, N-N_1)}{c(N_1, N-N_1)} = \frac{\frac{(N_1+1)!\,(N-N_1)!}{(N+2)!}}{\frac{N_1!\,(N-N_1)!}{(N+1)!}}$$

N_1+1

Substitute $c(m,k) = \dfrac{m!k!}{(m+k+1)!}$

Bayesian Inference

- Prior assumption about $\alpha : p(\alpha)$

  $\alpha \sim Uniform(0,1)$; $x \sim Bernoulli(\alpha)$
- Given evidence $\boldsymbol{X}$, what is the Expected A-Posteriori (EAP) $E_{\alpha}\left[\frac{p(\alpha)p(\boldsymbol{X}|\alpha)}{p(\boldsymbol{X})}\right] =$

$$= \frac{c(N_1+1, N-N_1)}{c(N_1, N-N_1)} = \frac{\frac{(N_1+1)!(N-N_1)!}{(N+2)!}}{\frac{N_1!(N-N_1)!}{(N+1)!}} = \frac{N_1+1}{N+2}$$

- Additive smoothing, add-1 smoothing

  Chance for sunrise tomorrow[Laplace]

Bayesian approach to overfit prevention

- Prior assumption about $\alpha : p(\alpha)$
- Given evidence $\boldsymbol{X}$, what is the Maximum A-Posteriori (MAP) $argmax_{\alpha}[p(\alpha)p(\boldsymbol{X}|\alpha)] =$
  $= argmax_{\alpha}[\log p(\alpha) + \log p(\boldsymbol{X}|\alpha)]$

Regression: Assuming θ is small

- Prior: $\Pr(\theta) \propto e^{-\frac{\lambda}{2}\|\theta\|^2}$

Assuming θ is small

- Prior: $\Pr(\theta) \propto e^{-\frac{\lambda}{2}||\theta||^2}$
- $\Pr(Data) = \Pr(Data|\theta) \times \Pr(\theta)$

- Posterior = Likelihood × Prior

Assuming θ is small

- Prior: $\Pr(\theta) \propto e^{-\frac{\lambda}{2}\|\theta\|^2}$
- $\Pr(Data) = \Pr(Data|\theta) \times \Pr(\theta)$

  $\log \Pr(Data) = \mathrm{l}(\theta) + \log \Pr(\theta)$
- Posterior = Likelihood × Prior

$\theta^* =$Max-aposteriori$=\operatorname{argmax}[\mathrm{l}(\theta) + \log \Pr(\theta)]$

Regularized Risk Minimization

- Empirical Risk Minimization gave overfitting & underfitting
- We want to add a penalty for using too many theta values

Regularized Risk Minimization

- Empirical Risk Minimization gave overfitting & underfitting
- We want to add a penalty for using too many theta values
- This gives us the Regularized Risk

$$R_{regularized}(\theta) = R_{empirical}(\theta) + Penalty(\theta)$$
$$= \frac{1}{N}\sum_{i=1}^{N} Loss\big(y_i, f(x_i;\theta)\big) + \frac{\lambda}{2}\|\theta\|^2$$

- Solution for Regularized Risk with Least Squares Loss:

Regularized Risk Minimization

- Empirical Risk Minimization gave overfitting & underfitting
- We want to add a penalty for using too many theta values
- This gives us the Regularized Risk

$$R_{regularized}(\theta) = R_{empirical}(\theta) + Penalty(\theta)$$
$$= \frac{1}{N}\sum_{i=1}^{N} Loss\big(y_i, f(x_i; \theta)\big) + \frac{\lambda}{2}\|\theta\|^2$$

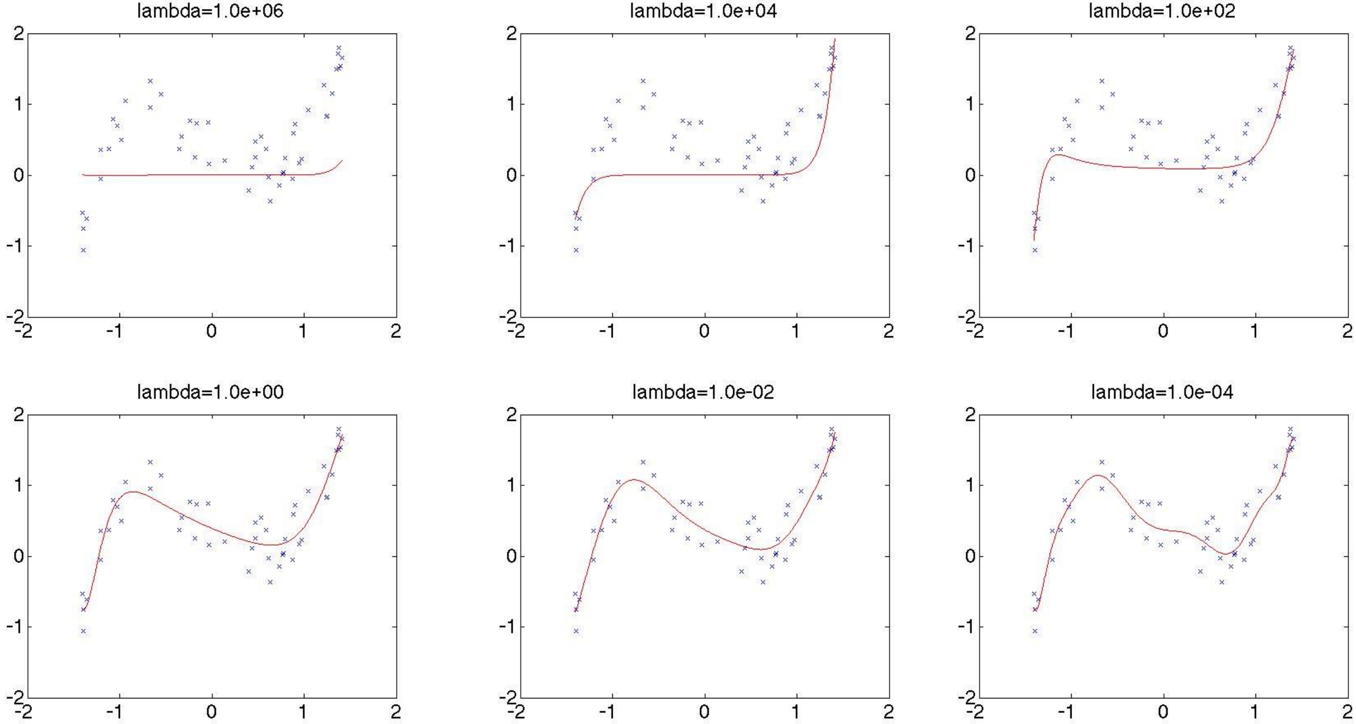
- Solution for Regularized Risk with Least Squares Loss:

$$\nabla_\theta R_{regularized} = 0$$
$$\nabla_\theta \left(\frac{1}{2N}\|\boldsymbol{y} - \boldsymbol{X}\theta\|^2 + \frac{\lambda}{2}\|\theta\|^2\right) = 0$$
$$\frac{1}{2N}(-2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\theta) + \frac{\lambda}{2}(2\theta) = 0$$

$$\theta^* = (\boldsymbol{X}^T\boldsymbol{X} + \lambda N I)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Regularized Risk Minimization

- Have D=16 features (or P=15 throughout)
- Try minimizing $R_{regularized}(\theta)$ to get θ^* with different λ
- Note that λ=0 give back Empirical Risk Minimization



Summary

- Inferring distribution parameters:
  - Max likelihood
  - Expected A-Posteriori
  - Maximum A-Posterior
- Regularization