Point Estimation Linear Regression

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

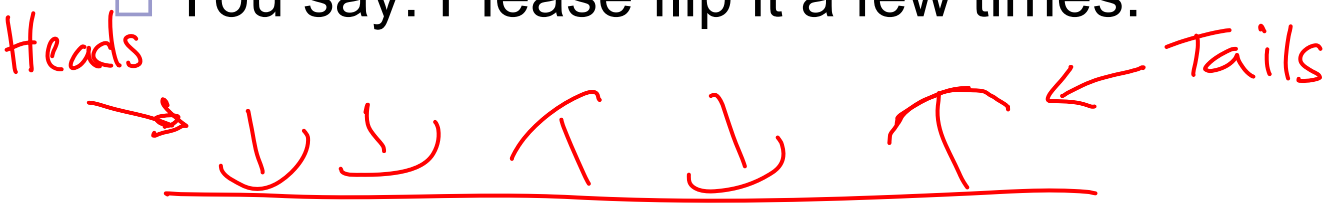
January 12th, 2005

Announcements

- Recitations – New Day and Room
  - Doherty Hall 1212
  - **Thursdays** – 5-6:30pm
  - Starting January 20th
- Use mailing list
  - 701-instructors@boysenberry.srv.cs.cmu.edu

Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
  - You say: Please flip it a few times:



  - You say: The probability is: $\theta = 3/5$
  - **He says: Why???**
  - You say: Because…

Thumbtack – Binomial Distribution

- P(Heads) = θ, P(Tails) = 1-θ

$P(HHTHT) = \theta\theta(1-\theta)\theta(1-\theta)$

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

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

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?

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

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

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

MLE:

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

simpler with ln

Your first learning algorithm

$\frac{d \ln x}{dx} = \frac{1}{x}$

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

$$= \arg\max_{\theta} \ \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

$$= \arg\max_{\theta} \ \alpha_H \ln \theta + \alpha_T \ln 1-\theta$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0}$

$$\frac{d}{d\theta} \ln P(D|\theta) = \frac{d}{d\theta} \alpha_H \ln \theta + \frac{d}{d\theta} \alpha_T \ln 1-\theta$$

$$= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0 \Rightarrow (1-\theta)\alpha_H = \alpha_T \theta$$

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

How many flips do I need?

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

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

Simple bound (based on Hoeffding's inequality)

- For $N = \alpha_H + \alpha_T$, and $\boxed{\hat{\theta} = \frac{\alpha_H}{\alpha_H + \alpha_T}}$

true

- Let θ^* be the true parameter, for any $\varepsilon > 0$: $\delta > 0$

0.1 1%

prob. of mistake how low

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

take ln of both sides:

$$\ln\left[2e^{-2N\varepsilon^2}\right] = \ln 2 - 2N\varepsilon^2 \leq \ln \delta$$

move things

$$\Rightarrow 2N\varepsilon^2 \geq \ln 2 + \ln 1/\delta \quad \Rightarrow \underbrace{N \geq \frac{1}{2\varepsilon^2}\left[\ln 2 + \ln 1/\delta\right]}$$

PAC: Probably Approx. Correct

PAC Learning

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the thumbtack parameter θ, within ε = 0.1, with probability at least 1-δ = 0.95. How many flips?

$$P(\mid \widehat{\theta} - \theta^* \mid \geq \epsilon) \;\leq\; 2e^{-2N\epsilon^2}$$

rearrange:

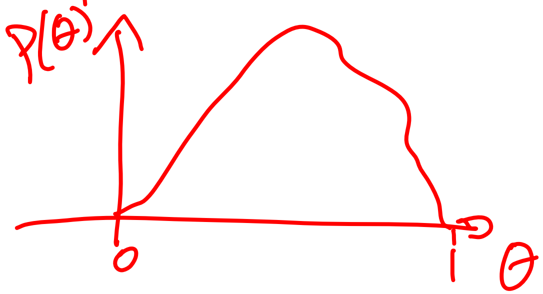
$$\boxed{N \geq \frac{1}{2\varepsilon^2}\left[\ln 2 + \ln 1/\delta\right]}$$

What about prior

- Billionaire says: Wait, I know that the thumbtack is “close” to 50-50. What can you?
- **You say: I can learn it the Bayesian way…**

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

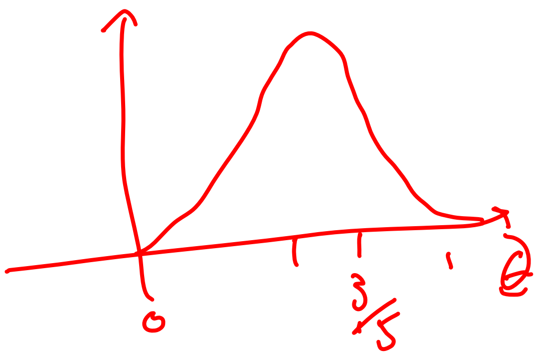
Expert prior



Collect data

H H T H T →

posterior



Bayesian Learning

- Use Bayes rule:

likelihood function

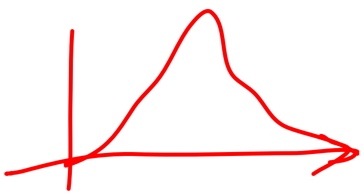
prior

$$\underbrace{P(\theta \mid \mathcal{D})}_{\text{posterior}} = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

normalization constant
⇒ make sure things add up to 1

- Or equivalently:

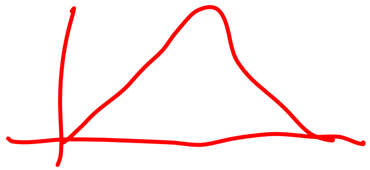
$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$



posterior

close 3/5

likelihood



prior

Bayesian Learning for Thumbtack

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$$

- Likelihood function is simply Binomial:

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

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

Beta prior distribution – P(θ)

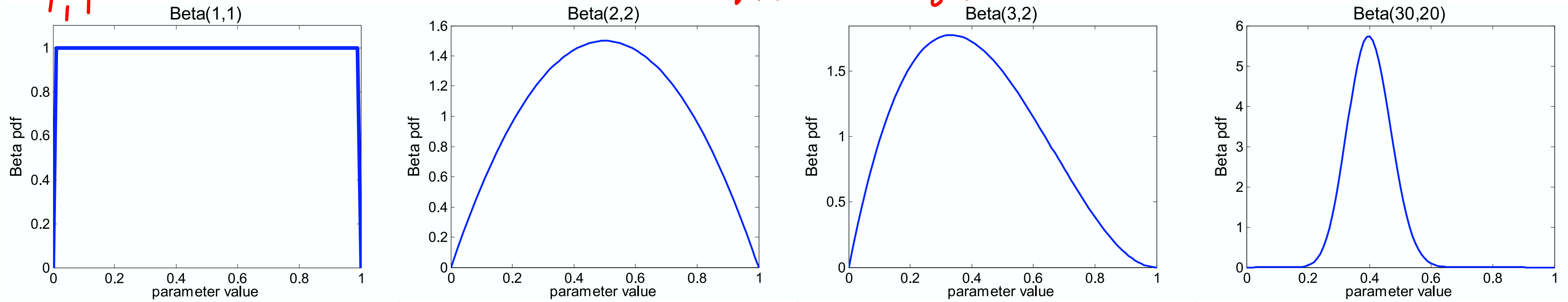
$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

function of θ

Normalization

1,1

30, 20



- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$

$$P(\theta \mid D) \propto \theta^{\alpha_H}(1-\theta)^{\alpha_T} \cdot \theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}$$

$$= \theta^{\alpha_H + \beta_H - 1}(1-\theta)^{\alpha_T + \beta_T - 1}$$

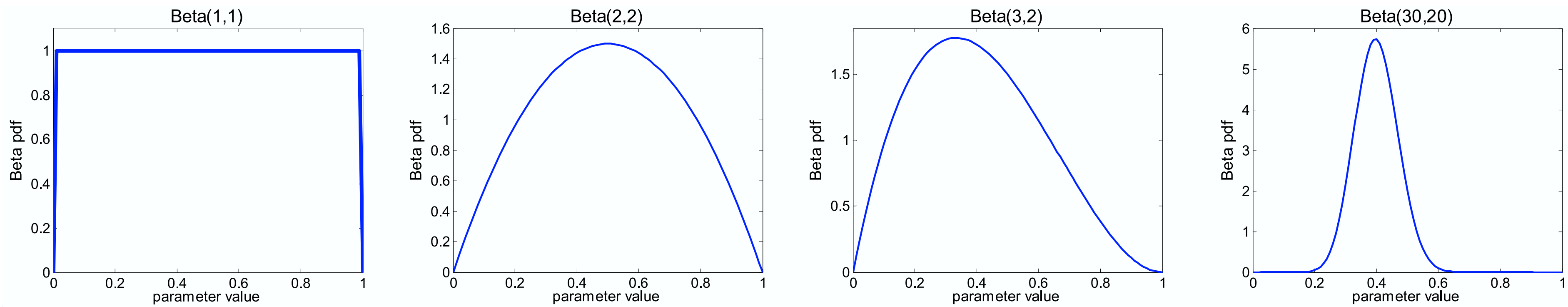
$$\sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

Posterior distribution

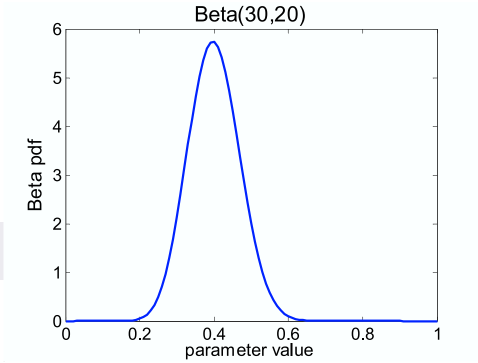
- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails

- Posterior distribution:

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



Using Bayesian posterior



- Posterior distribution:

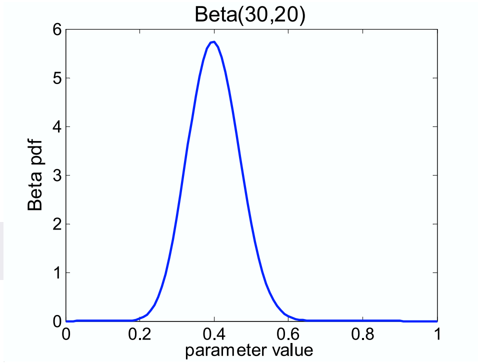
$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- Bayesian inference:
  - No longer single parameter:

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

  - Integral is often hard to compute

MAP: Maximum a posteriori approximation



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

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

approximation

- As more data is observed, Beta is more certain

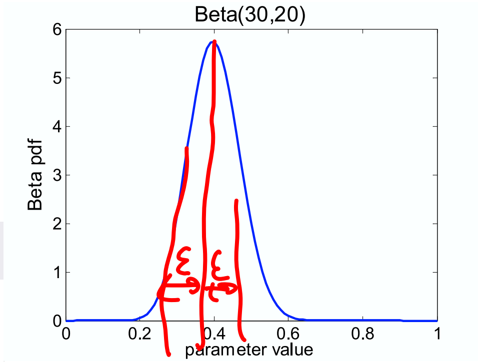
- MAP: use most likely parameter:

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

$$\stackrel{?}{=} \frac{\beta_H + \alpha_H}{\beta_H + \alpha_H + \beta_T + \alpha_T} \; ?$$

No!

MAP for Beta distribution



negative 1

$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H, \beta_T+\alpha_T)} \sim Beta(\beta_H+\alpha_H, \beta_T+\alpha_T)$$

- MAP: use most likely parameter:

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) = \boxed{\frac{\beta_H+\alpha_H-1}{\beta_H+\alpha_H+\beta_T+\alpha_T-2}}$$

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

What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- **You say: Let me tell you about Gaussians…**

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

MLE: estimate μ, σ
from iid samples

MLE for Gaussian

- Prob. of i.i.d. samples $x_1,\ldots,x_N$:

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

- Log-likelihood of data:

$$\begin{aligned}\ln P(\mathcal{D} \mid \mu, \sigma) &= \ln \overbrace{\left[\underbrace{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N}_{} \underbrace{\prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}}_{} \right]}^{} \\ &= \underbrace{-N \ln \sigma\sqrt{2\pi}}_{\text{log norm.}} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\end{aligned}$$

loglikeliho

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

$$
\begin{aligned}
\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) &= \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
&= \underbrace{\frac{d}{d\mu}\left[\cancel{-N \ln \sigma\sqrt{2\pi}}\right]}_{\text{doesn't depend on } \mu} \xrightarrow{} 0 - \sum_{i=1}^{N} \frac{d}{d\mu}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]
\end{aligned}
$$

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

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

MLE for variance

$$\frac{d}{d\sigma}\frac{1}{\sigma^2} = \frac{-2}{\sigma^3}$$

- Again, set derivative to zero:

$$\frac{d}{d\cancel{\mu}\sigma}\ln P(\mathcal{D}\mid\mu,\sigma) = \frac{d}{d\cancel{\mu}\sigma}\left[-N\ln\sigma\sqrt{2\pi} - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$-N(\ln\sigma + \ln\sqrt{2\pi})$$

$$= \frac{d}{d\cancel{\mu}\sigma}\left[\overbrace{-N\ln\sigma\sqrt{2\pi}}\right] - \sum_{i=1}^{N}\frac{d}{d\cancel{\mu}\sigma}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

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

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

Learning Gaussian parameters

- MLE:

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

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

- Bayesian learning is also possible
- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

Prediction of continuous variables

- Billionaire says: Wait, that's not what I meant!
- You says: Chill out, dude.
- He says: I want to predict a continuous variable for continuous inputs: I want to predict salaries from GPA.
- You say: **I can regress that…**

Salaries

$GPA \rightarrow \mathbb{R}$

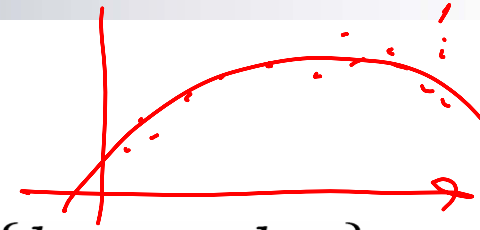
$S \approx aG + b$

The regression problem

$\langle 3.9, 90K \rangle$
$\langle 4.0, 112K \rangle$

input output

- **Instances:** $<\mathbf{x}_j, t_j>$
- **Learn:** Mapping from x to t($\mathbf{x}$)
- **Hypothesis space:**
  - Given, basis functions
  - Find coeffs $\boxed{\mathbf{w}=\{w_1,\ldots,w_k\}}$



ax^2+bx+c

x^2
x
1

$$H = \{h_1, \ldots, h_K\}$$

$$\underbrace{t(\mathbf{x})}_{\text{data}} \approx \widehat{f}(\mathbf{x}) = \sum_i w_i h_i(\mathbf{x})$$

- Precisely, minimize the residual error:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

- Solve with simple matrix operations:
  - Set derivative to zero
  - **Go to recitation Thursday 1/20**

- Data
- basis functions = features
- Matrix ops → w^*

But, why?

- Billionaire (again) says: Why sum squared error???
- You say: Gaussians, Dr. Gateson, Gaussians…
- Model:

Gaussian

$$\underbrace{P}_{\text{Gaussian}}(t \mid \mathbf{x}, \mathbf{w}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-\left[t - \overbrace{\sum_i w_i h_i(\mathbf{x})}^{\text{表 } \mu}\right]^2}{2\sigma^2}}$$

- Learn **w** using MLE

Maximizing log-likelihood

Maximize:

$$\boxed{\ln P(\mathcal{D} \mid \mathbf{w}, \sigma)} = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{j=1}^{N} e^{\frac{-[t_j - \sum_i w_i h_i(\mathbf{x}_j)]^2}{2\sigma^2}}$$

σ doesn't depend on w

$$= \cancel{\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)} N \leftarrow \sum_{j=1}^{N} \frac{\left[t_j - \sum_i w_i h_i(x_j)\right]^2}{2\sigma^2}$$

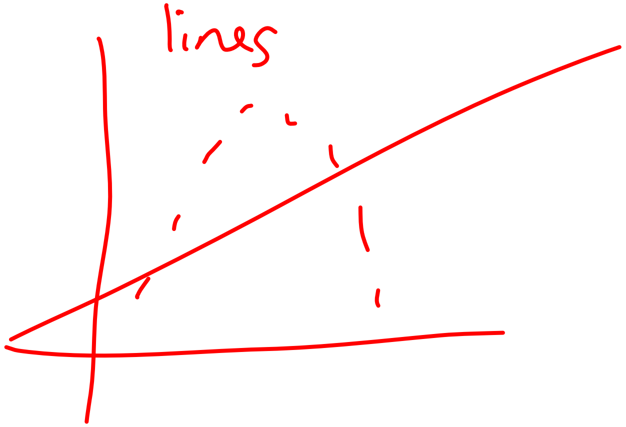
$$\equiv \text{minimizing } \sum_{j=1}^{N} \frac{\left[t_j - \sum_i w_i h_i(x_j)\right]^2}{2\sigma^2}$$

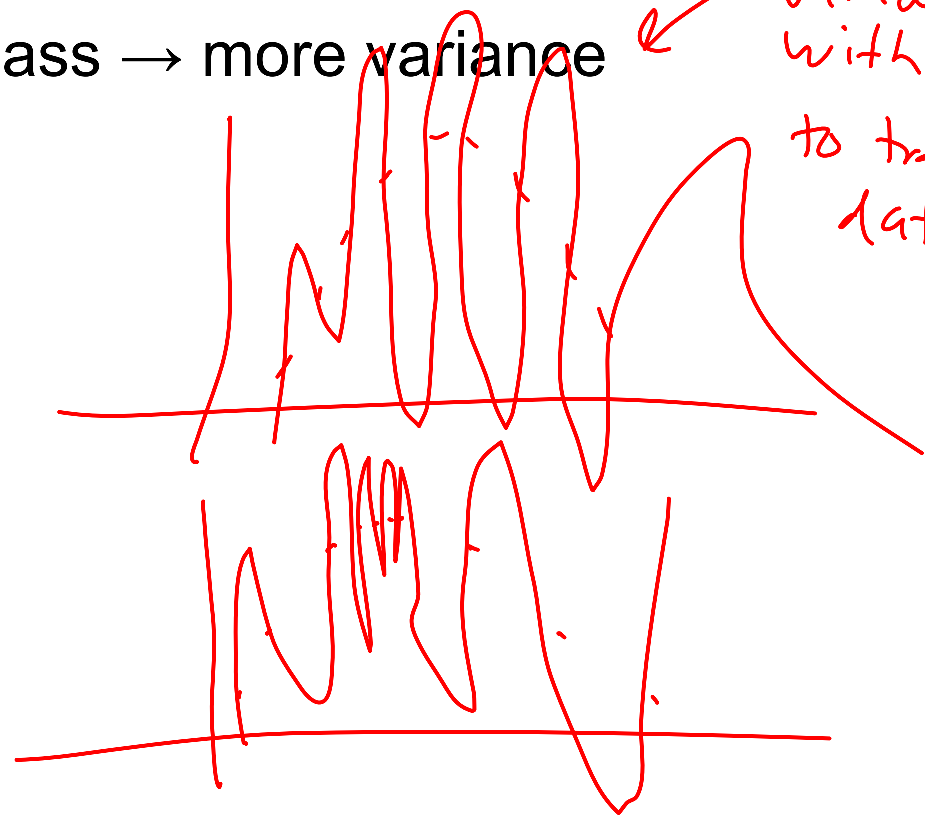
Bias-Variance Tradeoff

- Choice of hypothesis class introduces learning bias
  - More complex class → less bias
  - More complex class → more variance

expressivity of hypothesis class

variability with respect to training data





What you need to know

- Go to recitation for regression Thursday
  - And, other recitations too
- Point estimation:
  - MLE
  - Bayesian learning
  - MAP
- Gaussian estimation
- Regression
  - Basis function = features
  - Optimizing sum squared error
  - Relationship between regression and Gaussians
- Bias-Variance trade-off