TAM 598 Lecture 14:

Bayesian Linear Regression

Announcements:

- HW 4 covers lectures 13-16; due on Mar 31 (Mon)
- No class Monday Mar 24th

Today:

(1) Max Likelihood Estimation (MLE)

(2) Max A Posteriori Estimation (MPE)

(3) Bayesian Linear Regression

Least Squares Linear Regression – "traditional"

Can we interpret it from a more "modern" Bayesian or probalistic way?

observations $\underset{\sim}{X}_{1:n}$, outputs $y_{1:n}$

generalized linear model w/ m basis functions:

$$y(\underset{\sim}{x}; \underline{w}) = \sum_{j=1}^{m} w_j \phi_j(\underset{\sim}{x}) = \underline{w}^T \underline{\Phi}(\underset{\sim}{x})$$

(w_j: weights; ϕ_j: basis; $\underline{\Phi}$: design matrix)

$$\Phi_{ij} = \phi_j(\underset{\sim}{x}_i)$$

model the measurement process using a likelihood function

$$\underset{\sim}{y}_{1:n} \mid (\underset{\sim}{x}_{1:n}, \underline{w}) \sim p(\underline{y}_{1:n} \mid (\underset{\sim}{x}_{1:n}, \underline{w}))$$

assume that outcomes of single measurements are Gaussian, with mean $\underline{w}^T \underline{\Phi}(\underset{\sim}{x})$ and noise variance σ^2

$$p(y_i \mid \underset{\sim}{x}_i, \underline{w}, \sigma^2) = N(y_i \mid y(\underset{\sim}{x}_i, \underline{w}), \sigma^2) = N(y_i \mid \underline{w}^T \underline{\Phi}(\underset{\sim}{x}), \sigma^2)$$

notation: $\mathcal{N}(y \mid \mu, \sigma^2)$ is a pdf

$$p(y) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

for independent measurements, the likelihood of the data factorizes

$$p(\underline{y}_{1:n} \mid \underline{x}_{1:n}, \underline{w}) = \prod_{i=1}^{n} p(y_i \mid \underline{x}_i, \underline{w})$$

$$= \prod_{i=1}^{n} \mathcal{N}(y_i \mid \underline{w}^T \underline{\phi}(\underline{x}), \sigma^2)$$

$$\vdots$$

$n \times m$ design matrix

$$(*) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \| \underline{y}_{1:n} - \underline{\Phi}\, \underline{w} \|^2\right\}$$

Today:

(1) Max Likelihood Estimation (MLE)

- assume measurement outcomes are gaussian distributed
- find weights $\vec{w}$ and variance σ^2 that maximizes likelihood of measuring the observed data

(2) Max A Posteriori Estimation (MPE)

(3) Bayesian Linear Regression

(I) How do we find the parameters? Maximum likelihood for weights $\underline{w}$, and σ^2

$$\max_{\underline{w}, \sigma^2} \log p(\underline{y}_{1:n} \mid \underline{x}_{1:n}, \underline{w})$$

measurement

prediction

$$\max_{\underline{w}, \sigma^2} \left\{ -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2} \left\| \underline{y}_{1:n} - \underline{\underline{\Phi}}\,\underline{w} \right\|^2 \right\}$$

neg of sum of square errors

(*) So: maximizing likelihood w/r/t $\underline{w}$ is equivalent to minimizing sum of squares! And our weights should therefore satisfy

$$\underline{\underline{\Phi}}^T \underline{\underline{\Phi}}\, \underline{w} = \underline{\underline{\Phi}}^T \underline{y}_{1:n}$$

AND also we can estimate σ^2 as well via max likelihood

$$0 = \frac{d}{d\sigma^2}\{\ \} \Rightarrow$$

$$\sigma^2 = \frac{\left\| \underline{\underline{\Phi}}\,\underline{w} - \underline{y}_{1:n} \right\|^2}{n}$$

Now: we can incorporate uncertainty σ^2 when making predictions

point predictive distribution :

$$p(y \mid \underline{x}, \underline{w}, \sigma^2) = \mathcal{N}(y \mid \underline{w}^T \phi(\underline{x}), \sigma^2)$$

the measured output y is normal distributed around the model (least squares) prediction, with variance σ^2

Today:

(1) Max Likelihood Estimation (MLE)

(2) Max A Posteriori Estimation (MPE)

- ✓ - assume measurement outcomes are gaussian distributed as before
- ✓ - assume gaussian prior on weights $p(\underline{w})$ with zero mean, variance α
- ✓ - assume measurement uncertainty σ^2
- ✓ - find the weights $\underline{w}$ that maximize the prob. of the Bayesian posterior
- ✓ - helps to avoid overfitting

(3) Bayesian Linear Regression

II How do we find the parameters? Maximum a posteriori estimates

here: maximize the log prob of the posterior rather than the likelihood. Helps avoid overfitting

again: $y(\underline{x}; \underline{w}) = \underline{w}^T \underline{\underline{\phi}}(\underline{x})$

(*) now: assume

$$\underline{y}_{1:n} \mid \underline{\underline{x}}_{1:n}, \underline{w}, \sigma^2 \sim N\left(\underline{w}^T \underline{\underline{\phi}}(\underline{x}), \sigma^2\right)$$

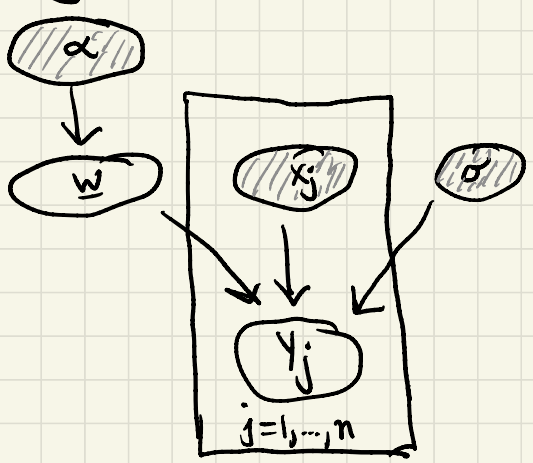
(*) new: uncertainty in model parameters described by a prior $\underline{w} \sim p(\underline{w})$

eg) gaussian prior on weights $p(\underline{w} \mid \alpha) = N\left(\underline{w} \mid \underline{0}, \alpha^{-1} \underline{\underline{I}}\right)$

mean values: $\underline{0}$; covariance matrix: $\alpha^{-1}\underline{\underline{I}}$

$$= \left(\frac{\alpha}{2\pi}\right)^{\frac{m}{2}} \exp\left\{-\frac{\alpha}{2}\|w\|^2\right\}$$

graphically



posterior from Bayes' Rule

$$p(\underline{w} \mid \underline{x}_{1:n}, \underline{y}_{1:n}, \sigma, \alpha) = \frac{p(\underline{y}_{1:n} \mid \underline{x}_{1:n}, \underline{w}, \sigma, \alpha)\, p(\underline{w} \mid \alpha)}{\int p(\underline{y}_{1:n} \mid \underline{x}_{1:n}, \underline{w}', \sigma, \alpha)\, p(\underline{w}' \mid \alpha)\, d\underline{w}'}$$

our state of knowledge about $\underline{w}$, after we see the data

Now assuming σ and α are known.

a point estimate of $\underline{w}$ is

$$\underline{w}_{MPE} = \underset{\underline{w}}{\operatorname{argmax}} \underbrace{p(\underline{y}_{1:n} | \underline{x}_{1:n}, \underline{w}, \sigma)}_{\text{likelihood}} \underbrace{p(\underline{w}|\alpha)}_{\text{prior}}$$

For gaussian likelihood and weights prior:

$$\log p(\underline{w} | \underline{x}_{1:n}, \underline{y}_{1:n}, \sigma, \alpha) = \left\{ -\frac{1}{2\sigma^2} \| \underline{\underline{\Phi}}\, \underline{w} - \underline{y}_{1:n} \|^2 - \frac{\alpha}{2} \|w\|^2 \right\}$$

maximizing:

$$\nabla_{\underline{w}} \{ \cdot \} = 0 \implies$$

$$\underline{w}_{MPE} = \frac{1}{\sigma^2} \left[\frac{1}{\sigma^2} \underline{\underline{\Phi}}^T \underline{\underline{\Phi}} + \alpha \underline{\underline{I}} \right]^{-1} \underline{\underline{\Phi}}^T \underline{y}_{1:n}$$

Today:

(1) Max Likelihood Estimation (MLE)

(2) Max A Posteriori Estimation (MPE)

(3) Bayesian Linear Regression

- just like (2) above, but now work with the full posterior distribution, not just a point estimate
- allows us to separate epistemic and aleatoric uncertainty

OR another way to see this: our posterior

$$p(\underline{w} \mid \underline{x}_{1:n}, \underline{y}_{1:n}, \sigma^2, \alpha) \propto \exp\left\{-\frac{1}{2\sigma^2}\|\underline{\underline{\Phi}}\,\underline{w} - \underline{y}_{1:n}\|^2 - \frac{\alpha}{2}\|w\|^2\right\}$$

algebra: rewrite in general form of a gaussian

$$\propto \exp\left\{-\frac{1}{2}(\underline{w} - \underline{m})^T \underline{\underline{S}}^{-1} (\underline{w} - \underline{m})\right\}$$

where $\underline{m} = \frac{1}{\sigma^2}\underline{\underline{S}}\,\underline{\underline{\Phi}}^T \underline{y}_{1:n}$ is the mean vector

$$\underline{\underline{S}} = \left[\frac{1}{\sigma^2}\underline{\underline{\Phi}}^T\underline{\underline{\Phi}} + \alpha\underline{\underline{I}}\right]^{-1}$$ is the covariance matrix

- If $\alpha \to 0$: OLS solution
- If $\sigma^2 \to 0$, data contribution dominates prior, more conf in $\underline{w}$

$\frac{1}{\sigma^2}\underline{\underline{\Phi}}^T\underline{\underline{\Phi}}$: from likelihood (data contrib)

$\alpha\underline{\underline{I}}$: from prior, regularizes est of $\underline{w}$

III How do we find the parameters? Bayesian Linear Regression

similar to II:

again: $y(\underline{x}; \underline{w}) = \underline{w}^T \underline{\phi}(\underline{x})$

$$\underline{y}_{1:n} \mid \underline{x}_{1:n}, \underline{w}, \sigma^2 \sim \mathcal{N}(\underline{w}^T \underline{\phi}(x), \sigma^2)$$

(μ, σ^2)

uncertainty in model parameters described by a prior $\underline{w} \sim p(\underline{w})$

W/ MPE

but no point estimate. Keep the posterior distribution and work with it directly. Why? Can now quantify the epistemic uncertainty arising from limited # of observations.

posterior is gaussian: $p(\underline{w} \mid \underline{x}_{1:n}, \underline{y}_{1:n}, \sigma, \alpha) = \mathcal{N}(\underline{w} \mid \underline{m}, \underline{\underline{S}})$

where $\underline{\underline{S}} = \left(\frac{1}{\sigma^2} \underline{\underline{\Phi}}^T \underline{\underline{\Phi}} + \alpha \underline{\underline{I}}\right)^{-1}$; $\underline{m} = \frac{1}{\sigma^2} \underline{\underline{S}} \underline{\underline{\Phi}}^T \underline{y}_{1:n}$

posterior predictive distribution: what can we say about y at some new $\underset{\sim}{x}$ after seeing the data?

$$p(y \mid \underset{\sim}{x}, \underset{\sim}{x}_{1:n}, \underset{\sim}{y}_{1:n}, \sigma, \alpha) = \int p(y \mid \underset{\sim}{x}, \underline{\underline{w}}, \sigma)\, p(\underline{\underline{w}} \mid \underset{\sim}{x}_{1:n}, \underset{\sim}{y}_{1:n}, \sigma, \alpha)\, dw$$

for all gaussian priors, this is analytically available

$$p(\underline{y} \mid \underset{\sim}{x}, \underset{\sim}{x}_{1:n}, \underset{\sim}{y}_{1:n}, \sigma, \alpha) = N(y \mid m(\underset{\sim}{x}), s^2(\underset{\sim}{x}))$$

where $m(\underset{\sim}{x}) = \underline{m}^T \phi(\underset{\sim}{x})$

$$s^2(\underset{\sim}{x}) = \underline{\phi}^T(\underset{\sim}{x})\, \underline{\underline{S}}\, \underline{\phi}(x) + \sigma^2$$

epistemic uncertainty

measurement noise