Machine learning II

Bayesian linear regression

Linear regression

Model for relationship between

- (several) independent variables $\mathbf{x} = (x_1, \ldots, x_{D-1})$
- and dependent variable y

$$y = w_0 + \sum_{i=1}^{D-1} w_i x_i + \epsilon$$

- *Structure*: Linear relationship with parameters $\mathbf{w}$
- *Noise*: Additive observation noise $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$

Linear regression

Model for relationship between

- (several) independent variables $\mathbf{x} = (x_1, \ldots, x_{D-1})$
- and dependent variable y

$$y = w_0 + \sum_{i=1}^{D-1} w_i x_i + \epsilon$$

- *Structure*: Linear relationship with parameters $\mathbf{w}$
- *Noise*: Additive observation noise $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$

Convenient matrix notation

$$y = \mathbf{w}^T \mathbf{x} + \epsilon$$

where $\mathbf{w}, \mathbf{x} \in \mathbb{R}^D$ and $x_0 \equiv 1$

Linear regression

Linear regression

- can be considered as a statistical model

$$p(y|\mathbf{x}) = \mathcal{N}(y|\mathbf{w}^T\mathbf{x}, \sigma_\epsilon)$$

- is easily generalized using *basis functions* $\Phi_i(\mathbf{x})$

$$y = \mathbf{w}^T\mathbf{\Phi}(\mathbf{x}) + \epsilon$$

Common basis functions include

- Polynomials of degree k, i.e. $\Phi_0(x) \equiv 1, \Phi_1(x) = x, \ldots, \Phi_k(x) = x^k$
- Radial basis functions, i.e. $\Phi_i(x) = e^{-\frac{(x-\mu_i)^2}{\sigma_i^2}}$

Linear regression

Example data $\{(\mathbf{x}_n, t_n)\}_{n=1}^{N}$

- Collect data in matrices

$$\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T)^T \in \mathbb{R}^{N \times D} \text{ and } \mathbf{t} = (t_1, \ldots, t_N)^T \in \mathbb{R}^{N \times 1}$$

  $\mathbf{X}$ is also called *design matrix*

- Likelihood $p(Data|\theta)$

$$\begin{aligned} p(\mathbf{t}|\mathbf{X}, \theta) &= \prod_{n=1}^{N} \mathcal{N}(t_n | \mathbf{w}^T \mathbf{x}_n, \sigma_\epsilon) \\ &= (2\pi\sigma_\epsilon^2)^{-\frac{N}{2}} e^{-\frac{1}{2}\sum_{n=1}^{N} \frac{(t_n - \mathbf{w}^T \mathbf{x}_n)^2}{\sigma_\epsilon^2}} \end{aligned}$$

  with parameters $\theta = (\mathbf{w}, \sigma_\epsilon)$
  - weights $\mathbf{w} = (w_0, \ldots, w_D)^T$
  - noise variance σ_ϵ^2

Ordinary least squares

Maximum likelihood solution:

$$\max_{\mathbf{w},\sigma_\epsilon} \ln p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma_\epsilon) = \max_{\mathbf{w},\sigma} -\frac{N}{2}\ln\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2$$

- Find optimal weight vector $\mathbf{w}^*$:

$$\begin{aligned}
\frac{\partial}{\partial w_i}\ln p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma_\epsilon) &= \frac{1}{\sigma_\epsilon^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)x_{in} \overset{!}{=} 0 \\
\implies \mathbf{w}^{*T}\left(\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T\right) &= \sum_{i=1}^{N} t_n\mathbf{x}_n^T \\
\mathbf{w}^* &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}
\end{aligned}$$

- Minimizes the squared error $\frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2$
- $\mathbf{X}^\dagger = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is *pseudo-inverse* of design matrix $\mathbf{X}$

Ordinary least squares

Maximum likelihood solution:

$$\max_{\mathbf{w},\sigma_\epsilon} \ln p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma_\epsilon) = \max_{\mathbf{w},\sigma} -\frac{N}{2}\ln\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2$$

- Find optimal weight vector $\mathbf{w}^*$:

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

- Find optimal noise precision τ_ϵ^*:

$$\begin{aligned}
\frac{\partial}{\partial\tau_\epsilon}\ln p(\mathbf{t}|\mathbf{X},\mathbf{w}^*,\tau_\epsilon) &= \frac{N}{2}\frac{1}{\tau_\epsilon} - \frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2 \overset{!}{=} 0 \\
\implies \frac{1}{\tau_\epsilon^*} &= \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2
\end{aligned}$$

Bayesian linear regression

- Assume that noise precision τ_ϵ is known
- Conjugate prior for weights $\mathbf{w}$ is Gaussian

$$p(\mathbf{w}|\mathbf{0}, \Sigma_0)$$

Posterior is again Gaussian with

- covariance matrix

$$\Sigma_N = (\Sigma_0^{-1} + \tau_\epsilon \mathbf{X}^T \mathbf{X})^{-1}$$

- mean

$$\mu_N = \tau_\epsilon \Sigma_N \mathbf{X}^T \mathbf{t}$$

Bayesian linear regression

Run demo . . .

Predictive Distribution

- Predict new data point at $\mathbf{x}_{new}$
- Predictive distribution:

$$p(t_{new}|\mathbf{x}_{new}) = \int p(t_{new}|\mathbf{x}_{new}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}, \mathbf{t}) d\mathbf{w}$$

Again Gaussian distribution with

- mean $\mu_N^T \mathbf{x}_{new}$
- variance $\sigma_\epsilon^2 + \mathbf{x}_{new}^T \Sigma_N \mathbf{x}_{new}$

Noise variance + uncertainty about parameters

Do not pick "best" parameters, but take uncertainty into account $\implies$ Bayesian slogan:

Don't optimize, integrate!

Regularization

- Compare ML solution

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

  - Minimizes $\frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2$
- and posterior maximum ($\Sigma_0^{-1} = \tau_0\mathbf{I}$)

$$\mu_N = \tau_\epsilon(\tau_0\mathbf{I} + \tau_\epsilon\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

  - Minimizes regularized mean squared error

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

where $\lambda = \frac{\tau_0}{\tau_\epsilon}$

squared error + regularizer

Bayesian model selection

- Maximum likelihood $\max_\theta p(\mathbf{t}|\mathbf{X}, \theta)$ improves when number of parameters increases.
  This leads to *over-fitting* as model picks up noise in data.
- To achieve low expected loss we need to control model complexity
  Regularization favors less flexible models by penalizing large weights
- **Marginal likelihood/Evidence**:

$$p(\mathbf{t}|\mathbf{X}) = \int p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w}$$

  - Bayesian answer to over-fitting
  - Probability of data accounting for parameter uncertainty

Bayesian model selection

Consider set of models $\{\mathcal{M}_i\}$, e.g. regression models with different number/subset of covariates:

- Prior probability of each model $p(\mathcal{M}_i)$
- Compute posterior:

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)$$

  Note that each model has parameters $\mathbf{w}_i$ and thus

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}, \mathbf{w}_i|\mathcal{M}_i)d\mathbf{w}_i = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}_i|\mathcal{M}_i)d\mathbf{w}_i$$

  Marginal likelihood (wrt parameters) is likelihood (wrt model)!
- Note: Marginal likelihood requires proper prior

Bayesian model selection

Two models $\mathcal{M}_1, \mathcal{M}_2$ are then compared based on posterior odds

$$\frac{p(\mathcal{M}_1|\mathcal{D})}{p(\mathcal{M}_2|\mathcal{D})} = \underbrace{\frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)}}_{\text{Bayes factor}} \cdot \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{\text{Prior odds}}$$

Interpretation of evidence from Bayes factors as suggested by Jeffreys:

Bayes factor	**decibans**	**Evidence**
< 1	< 0	negative (supports $\mathcal{M}_2$)
$1 - 10^{\frac{1}{2}}$	$0 - 5$	barely worth mentioning
$10^{\frac{1}{2}} - 10$	$5 - 10$	substantial
$10 - 10^{\frac{3}{2}}$	$10 - 15$	strong
$10^{\frac{3}{2}} - 100$	$15 - 20$	very strong
> 100	> 20	decisive

Bayesian model selection

Marginal likelihood can be computed analytically:

- Prior: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \tau_0 \mathbf{I})$
- Likelihood: $p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{t}|\mathbf{X}\mathbf{w}, \tau_\epsilon \mathbf{I})$
- Evidence:

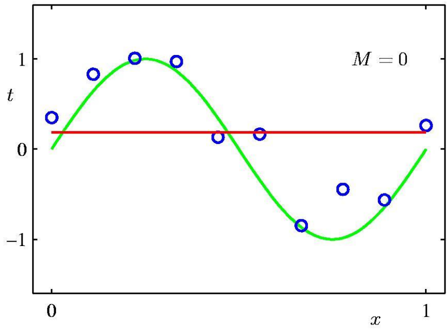
$$
\begin{aligned}
\ln p(\mathbf{t}|\mathbf{X}) \quad &= \quad \ln \frac{p(\mathbf{t}|\mathbf{X}, \mathbf{w}^*)p(\mathbf{w})^*}{p(\mathbf{w}^*|\mathbf{X}, \mathbf{t})} \text{ for any } \mathbf{w}^* \\
&\underbrace{=}_{\mathbf{w}^*=\mu_N} \quad \frac{D}{2}\ln \tau_0 + \frac{N}{2}\ln \tau_\epsilon - \frac{N}{2}\ln 2\pi \\
&- \quad \frac{\tau_\epsilon}{2}(\mathbf{X}\mu_N - \mathbf{t})^T(\mathbf{X}\mu_N - \mathbf{t}) - \frac{\tau_0}{2}\mu_N^T\mu_N - \frac{1}{2}\ln|\mathbf{A}|
\end{aligned}
$$

where

$$
\begin{aligned}
\mu_{\mathbf{N}} &= \tau_\epsilon \mathbf{A}^{-1}\mathbf{X}^T\mathbf{t} \\
\mathbf{A} &= \tau_0 \mathbf{I} + \tau_\epsilon \mathbf{X}^T\mathbf{X} = \Sigma_N^{-1}
\end{aligned}
$$

Bayesian model selection

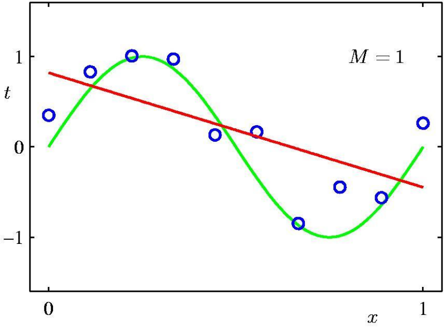
0^{th} Order Polynomial



Slide from `https://www.microsoft.com/en-us/research/people/cmbishop/`

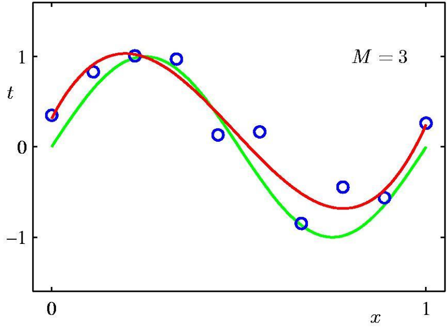
Bayesian model selection

1st Order Polynomial



Slide from `https://www.microsoft.com/en-us/research/people/cmbishop/`

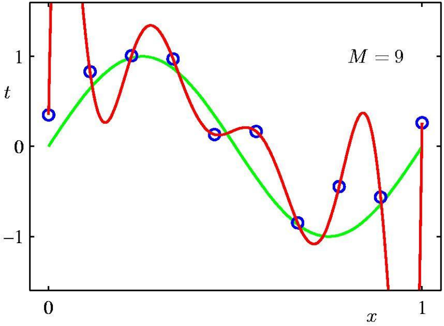
3^{rd} Order Polynomial



Slide from `https://www.microsoft.com/en-us/research/people/cmbishop/`

Bayesian model selection

9th Order Polynomial

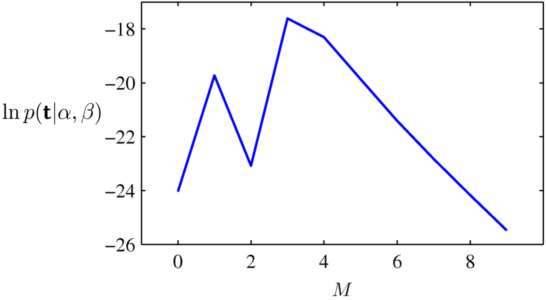


Slide from https://www.microsoft.com/en-us/research/people/cmbishop/

The Evidence Approximation (3)

Example: sinusoidal data, M^{th} degree polynomial,

$\alpha = 5 \times 10^{-3}$



Slide from https://www.microsoft.com/en-us/research/people/cmbishop/

Evidence approximation

How to choose hyperparameters, e.g. τ_0?

- Bayesian ideal: Assume prior $p(\tau_0)$ and integrate
  - Often analytically intractable
  - Introduces new hyperparameters ... have to stop at some point
- Evidence approximation:
  - Optimize marginal likelihood $p(\mathcal{D}|\tau_0)$, i.e. *ML II*
  - Tends to work well in practice with little over-fitting
  - Model selection from $\{\mathcal{M}_{\tau_0}\}_{\tau_0>0}$

Cross-validation

- Idea: Select model based on generalization error, i.e. best predictions on novel data
- *Cross-validation* estimates generalization error:
  - Partition data set D into K parts $D_1, \ldots, D_K$
  - For each $i = 1, \ldots, K$
    1. Train model on $D \setminus D_i$
    2. Evaluate model on D_i, e.g. using predictive distribution

    Predictive performance is estimated by average prediction error across $D_1, \ldots, D_K$
- Common choices in practice
  - $K = 10$: Good compromise between bias (due to small training set) and variance (due to small test sets)
  - $K = N$: Leaving-one-out cross-validation LOOCV
    Often efficient if data point can easily be "removed" from trained model

Model selection

Model selection summary:

	Evidence	**Cross-validation**
Philosophy	Bayesian	Frequentist
	Prior $p(\theta)$ matters	Prior insensitive
	Model all data $p(D)$	Predict part of data $p(D_i \mid D \setminus D_i)$
	Often computationally demanding	
Evaluation	Probability	Different error functions
Asymptotics	Consistent	Often inconsistent, e.g. LOOCV
	$BIC = -2 \log p(D \mid \theta_{ML}) + D \log N$	$AIC = -2 \log p(D \mid \theta_{ML}) + 2D$
Interpretation	Data compression	Prediction