Advanced Statistical Inference
Bayesian Linear Regression

Maurizio Filippone
Maurizio.Filippone@eurecom.fr

Department of Data Science
EURECOM

Recap - Probabilities

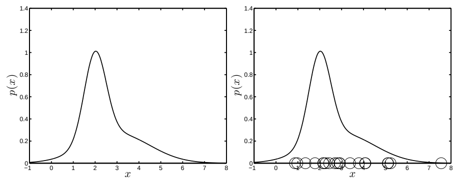
Consider two continuous random variables x and y

- Sum rule:
$$p(x) = \int p(x, y) dy$$
- Product rule:
$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$
- Bayes' rule:
$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$
- NOTE: Bayes' rule is a direct consequence of the product rule

Recap - Expectations

- Consider a random variable with density $p(x)$
- Imagine wanting to know the average value of x, $\tilde{x}$.
- Generate S samples, $x_1, \ldots, x_S$
- Average the samples:

$$\tilde{x} \approx \frac{1}{S} \sum_{s=1}^{S} x_s$$



Recap - Expectations

- Our sample based approximation to $\tilde{x}$ will get better as we take more samples.
- We can also (sometimes) compute it exactly using **expectations**.
  - Discrete:
  $$\tilde{x} = \mathrm{E}_{p(x)}(x) = \sum_{x} x \, p(x)$$
  - Continuous:
  $$\tilde{x} = \mathrm{E}_{p(x)}(x) = \int x \, p(x) \, dx$$

Recap - Expectations

- Example:
  - X is outcome of rolling die. $P(X = x) = 1/6$
$$\tilde{x} = \sum_x x\, P(X = x) = 3.5$$
  - X is uniform distributed RV between a and b
$$\tilde{x} = \int_{x=a}^{x=b} x p(x)\ dx = (b - a)/2$$

Expectations

- In general:
$$\mathrm{E}_{p(x)}[f(x)] = \int f(x)\, p(x)\, dx$$
- Some important properties:
$$\mathrm{E}_{p(x)}[f(x)] \neq f\left(\mathrm{E}_{p(x)}[x]\right)$$
$$\mathrm{E}_{p(x)}[k\, f(x)] = k\, \mathrm{E}_{p(x)}[f(x)]$$
- Mean and variance
$$\mu = \mathrm{E}_{p(x)}[x]$$
$$\sigma^2 = \mathrm{E}_{p(x)}[(x - \mu)^2] = \mathrm{E}_{p(x)}[x^2] - \mu^2$$

Expectations

- In general:
$$\mathrm{E}_{p(x)}[f(x)] = \int f(x)\, p(x)\, dx$$
- For vectors of random variables:
$$\mathrm{E}_{p(\mathbf{x})}[f(\mathbf{x})] = \int f(\mathbf{x})\, p(\mathbf{x})\, dx$$
  - Mean and covariance:
$$\boldsymbol{\mu} = \mathrm{E}_{p(\mathbf{x})}[\mathbf{x}]$$
$$\begin{aligned} \mathrm{cov}(x) &= \mathrm{E}_{p(\mathbf{x})}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] \\ &= \mathrm{E}_{p(\mathbf{x})}[\mathbf{x}\mathbf{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top \end{aligned}$$

The Gaussian Distribution

Consider a continuous random variable v

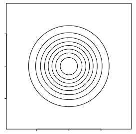
- The Gaussian probability density function is:
$$p(v|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(v - \mu)^2\right\}$$
- μ is the mean
- σ^2 is the variance

The Multivariate Gaussian Distribution

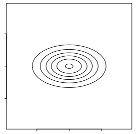
- Consider $\mathbf{v} = (v_1, \ldots, v_D)^\top$ with joint Gaussian distribution

$$p(\mathbf{v}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{v}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

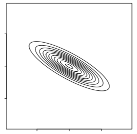
$$\mathcal{N}(\mathbf{v}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{v}-\boldsymbol{\mu})\right\}$$



$$\boldsymbol{\Sigma} = \begin{bmatrix} 9 & 0 \\ 0 & 9 \end{bmatrix}$$



$$\boldsymbol{\Sigma} = \begin{bmatrix} 9 & 0 \\ 0 & 3 \end{bmatrix}$$



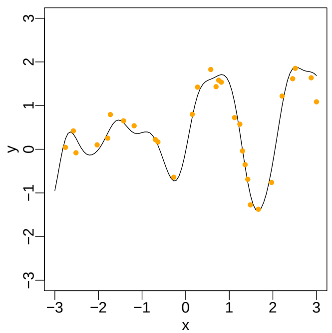
$$\boldsymbol{\Sigma} = \begin{bmatrix} 8 & -4 \\ -4 & 3 \end{bmatrix}$$

Expectations – Gaussians

- Univariate
  - $p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2)$
  - Mean: $\mathrm{E}_{p(x)}[x] = \mu$
  - Variance: $\mathrm{E}_{p(x)}\left[(x-\mu)^2\right] = \sigma^2$
- Multivariate
  - $p(\mathbf{x}|\mu, \sigma^2) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Mean: $\mathrm{E}_{p(\mathbf{x})}[\mathbf{x}] = \boldsymbol{\mu}$
  - Variance: $\mathrm{E}_{p(\mathbf{x})}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right] = \boldsymbol{\Sigma}$

Working example

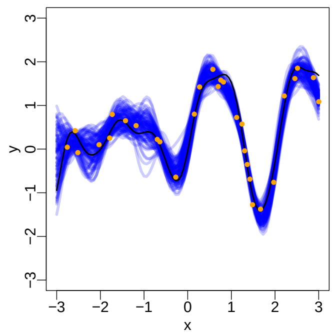
- Some data



- In this course we will learn ways to estimate functions that interpolate data...

Working example

- Function estimation...



- ... with confidence intervals
- Useful for **uncertainty** quantification

Definitions

- Features, inputs, covariates, or attributes $\mathbf{x}$:

$$\mathbf{x} \in \mathbb{R}^D$$

$$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^\top$$

- Labels, outputs, or responses:

$$\mathbf{y} \in \mathbb{R}^O$$

$$\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)^\top$$

Linear Regression - Definitions

- Data is a set of N pairs feature vectors and labels:

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,\dots,N}$$

- GOAL: Estimate a function

$$\mathbf{f}(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}^O$$

- For simplicity, we will assume $O = 1$ (univariate labels)

$$\mathbf{y} = (y_1, \dots, y_N)^\top$$

so we aim to estimate:

$$f(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}$$

Linear Models for Regression

- Implement a linear combination of basis functions

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i=1}^{D} w_i \varphi_i(\mathbf{x}) \\ &= \mathbf{w}^\top \boldsymbol{\varphi}(\mathbf{x}) \end{aligned}$$

with

$$\boldsymbol{\varphi}(\mathbf{x}) = (\varphi_1(\mathbf{x}), \dots, \varphi_D(\mathbf{x}))^\top$$

Linear Models for Regression

- For simplicity we will start with linear functions

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i=1}^{D} w_i x_i \\ &= \mathbf{w}^\top \mathbf{x} \end{aligned}$$

Linear Regression as Loss Minimization

- Definition of the quadratic loss function:

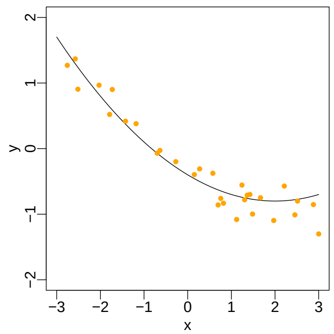
$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{N} [y_i - \mathbf{w}^\top \mathbf{x}_i]^2 \\ &= \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2\end{aligned}$$

- Solution to the regression problem is:

$$\nabla_{\mathbf{w}} \mathcal{L} = \mathbf{0} \quad \Longrightarrow \quad \widehat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Working example

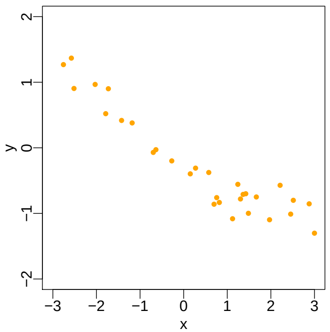
- Some data generated from a known function



Bayesian Linear Regression
└Loss Minimization in Linear Regression
└Loss minimization

Working example

- Some data generated from a known function



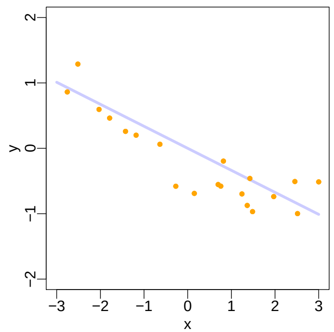
- In reality we only observe data and we want to estimate the generating function

Bayesian Linear Regression
└Loss Minimization in Linear Regression
└Loss minimization

Working example

- Solution obtained when optimizing the loss

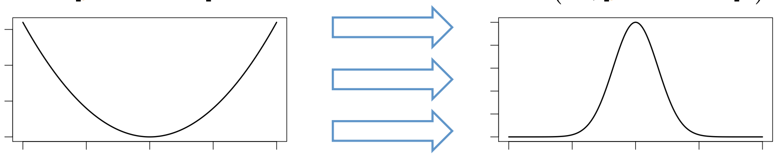
$$f(\mathbf{x}) = \widehat{\mathbf{w}}^\top \mathbf{x}$$



Probabilistic Interpretation of Loss Minimization

- Consider a simple transformation of the loss function

$[y_i - \mathbf{w}^\top \mathbf{x}_i]^2$ $\qquad$ $\exp\left(-\gamma[y_i - \mathbf{w}^\top \mathbf{x}_i]^2\right)$



- Minimizing the quadratic loss equivalent to maximizing the Gaussian likelihood function

$$\begin{aligned}\exp\left(-\gamma \mathcal{L}\right) &= \exp\left(-\gamma \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2\right) \\ &\propto \mathcal{N}\left(\mathbf{y} | \mathbf{X}\mathbf{w}, \frac{1}{2\gamma}\right) \qquad \text{Gaussian distribution}\end{aligned}$$

Probabilistic Interpretation of Loss Minimization

- The likelihood $\mathcal{N}\left(\mathbf{y} | \mathbf{X}\mathbf{w}, \frac{1}{2\gamma}\right)$ hints to the fact that we are assuming:

$$y_i = \mathbf{w}^\top \mathbf{x}_i + \varepsilon_i$$

  with $\varepsilon_i \sim \mathcal{N}(\varepsilon_i | 0, \sigma^2 = \frac{1}{2\gamma})$
- In vectorial form:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$$

  with $\boldsymbol{\epsilon} \sim \mathcal{N}(\varepsilon | 0, \sigma^2 \mathbf{I})$
- Remark: the likelihood is not a probability!

Probabilistic Interpretation of Loss Minimization

- Recall that the Maximum-Likelihood solution is

$$\widehat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- Now we can also maximize the log-likelihood to obtain the optimal σ^2:

$$\frac{\partial \log[p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \sigma^2)]}{\partial \sigma^2} = 0$$

  yielding

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{y} - \mathbf{X}\widehat{\mathbf{w}})^\top (\mathbf{y} - \mathbf{X}\widehat{\mathbf{w}})$$

Properties of the Maximum-Likelihood Estimator

- Are there any useful properties for the estimator $\widehat{\mathbf{w}}$?
- The estimator $\widehat{\mathbf{w}}$ is **unbiased**, that is:

$$\mathrm{E}_{p(\mathbf{y} | \mathbf{X}, \mathbf{w})}[\widehat{\mathbf{w}}] = \int \widehat{\mathbf{w}}\, p(\mathbf{y} | \mathbf{X}, \mathbf{w})\, d\mathbf{y} = \mathbf{w}$$

Properties of the Maximum-Likelihood Estimator

- The proof is rather simple:

$$\begin{aligned} \mathrm{E}_{p(\mathbf{y}|\mathbf{X},\mathbf{w})}[\widehat{\mathbf{w}}] &= \mathrm{E}_{p(\mathbf{y}|\mathbf{X},\mathbf{w})}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}] \\ &= \int (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\,\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})\,d\mathbf{y} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \int \mathbf{y}\,\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})\,d\mathbf{y} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\,\mathbf{w} \\ &= \mathbf{w} \end{aligned} \tag{1}$$

Properties of the Maximum-Likelihood Estimator

- The estimate of the optimal σ^2 is biased!

$$\begin{aligned} \mathrm{E}_{p(\mathbf{y}|\mathbf{X},\mathbf{w})}\left(\widehat{\sigma^2}\right) &= \frac{1}{N}\mathrm{E}_{p(\mathbf{y}|\mathbf{X},\mathbf{w})}\left[(\mathbf{y}-\mathbf{X}\widehat{\mathbf{w}})^\top(\mathbf{y}-\mathbf{X}\widehat{\mathbf{w}})\right] \\ &= \sigma^2\left(1-\frac{D}{N}\right) \end{aligned}$$

Properties of the Maximum-Likelihood Estimator

- The proof uses these two useful identities:

Expectation of quadratic form for Gaussian variables

$$\begin{aligned} p(\mathbf{v}) &= \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ \mathrm{E}_{p(\mathbf{v})}\left(\mathbf{v}^\top\mathbf{A}\mathbf{v}\right) &= \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu} \\ \mathrm{Tr}(\mathbf{A}) &= \sum_i \mathbf{A}_{ii} \end{aligned}$$

Permutation invariance of the trace operator

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$$

Properties of the Maximum-Likelihood Estimator

$$\begin{aligned} \mathrm{E}_{p(\mathbf{y}|\mathbf{X},\mathbf{w},\sigma^2)}(\widehat{\sigma^2}) &= \frac{1}{N}(\mathrm{Tr}(\sigma^2\mathbf{I}) + \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}) \\ &\quad -\frac{1}{N}(\mathrm{Tr}(\sigma^2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) + \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}) \\ &= \sigma^2 - \frac{\sigma^2}{N}\mathrm{Tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) \\ &= \sigma^2 - \frac{\sigma^2}{N}\mathrm{Tr}(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}) \\ &= \sigma^2\left(1-\frac{D}{N}\right) \end{aligned}$$

Where D is the dimensionality of $\mathbf{w}$.

Model Selection

- How can we prefer one model over another?
- Lowest loss / highest likelihood?
- **NO!**
- Higher model complexity yields lower loss / higher likelihood...
- ...but it usually does not generalize well on test data.

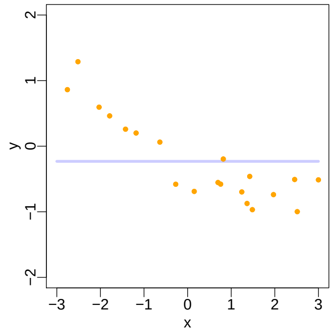
Bayesian Linear Regression
└Loss Minimization in Linear Regression
└Model Selection

Model Selection - Effect of increasing model complexity

- Consider polynomial functions:

$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 0$



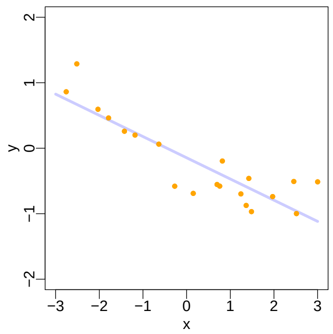
Bayesian Linear Regression
└Loss Minimization in Linear Regression
└Model Selection

Model Selection - Effect of increasing model complexity

- Consider polynomial functions:

$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 1$



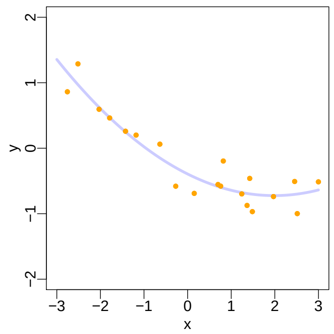
Bayesian Linear Regression
└Loss Minimization in Linear Regression
└Model Selection

Model Selection - Effect of increasing model complexity

- Consider polynomial functions:

$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 2$

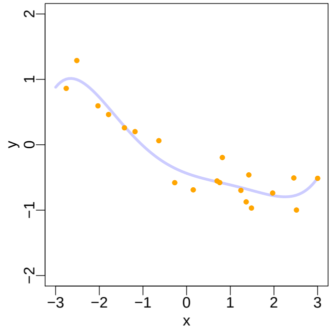


Model Selection - Effect of increasing model complexity

- Consider polynomial functions:

$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 5$

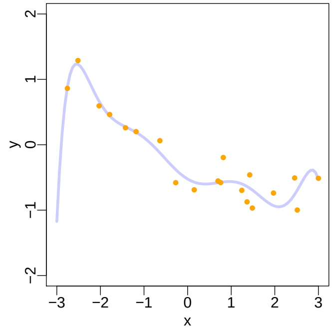


Model Selection - Effect of increasing model complexity

- Consider polynomial functions:

$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 8$

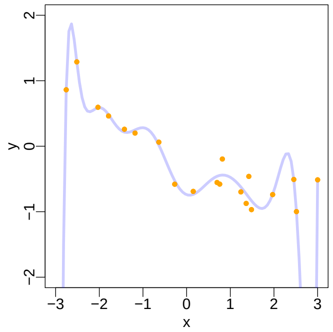


Model Selection - Effect of increasing model complexity

- Consider polynomial functions:

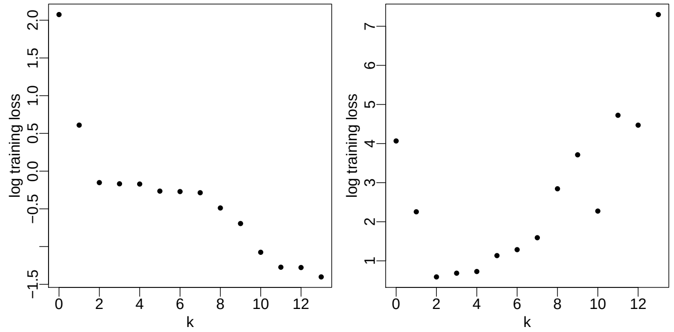
$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 13$



Model Selection - Effect of increasing model complexity

- Training loss decreases with k but test loss increases



Validation on "unseen" data

- Cross-validation is a safe way to do model selection
- Predictions evaluated using validation loss:

$$\mathcal{L}_v = \frac{1}{N_{\text{test}}} \sum_{i \in \mathcal{I}_{\text{test}}} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2$$

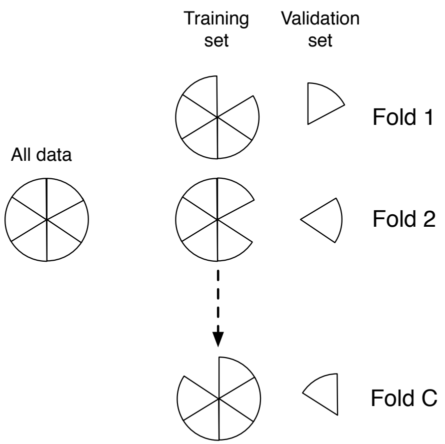
- Or the validation log-likelihood:

$$\log[p(\mathbf{y}_{\text{test}} | \mathbf{X}_{\text{test}}, \widehat{\mathbf{w}}, \sigma^2)] = -\frac{1}{2\sigma^2} \sum_{i \in \mathcal{I}_{\text{test}}} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2$$

How should we choose which data to hold back?

- In some applications it will be clear
- In many cases – pick it randomly
- Do it more than once – average the results
- Do cross-validation
  - Split the data into C equal sets. Train on $C - 1$, test on remaining.

Cross-validation



Average performance over the C 'folds'.

Leave-one-out Cross-validation

- Cross-validation can be repeated to make results more accurate
- e.g. Doing 10-fold CV 10 times gives us 100 performance values to average over
- Extreme example is when $C = N$ so each fold includes one input-label pair
  - Leave-one-out (LOO) CV

Computational issues

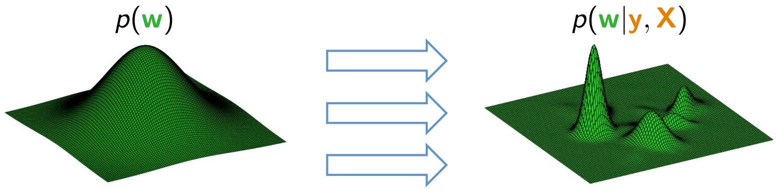
- CV and LOOCV let us choose from a set of models based on predictive performance.
- This comes at a computational cost:
  - For C-fold CV, need to train our model C times.
  - For LOO-CV, need to train out model N times.
- For $y = \mathbf{w}^\top \mathbf{x}$, this is feasible if K (number of terms in function) isn't too big:

$$\begin{aligned} y &= \sum_{k=0}^{K} w_k x_k \\ \widehat{\mathbf{w}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \end{aligned}$$

- For some models we need to use $C \ll N$.

Bayesian Inference

- Inputs : $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^\top$
- Labels : $\mathbf{y} = (y_1, \dots, y_N)^\top$
- Weights : $\mathbf{w} = (w_1, \dots, w_D)^\top$



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

Bayesian Linear Regression

- Modeling observations as noisy realizations of a linear combination of the features:

$$p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I})$$

- Gaussian prior over model parameters:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{S})$$

Bayesian Linear Regression

- Bayes rule:

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

- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$
  - Distribution over parameters after observing data
- **Likelihood** : $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
  - Measure of "fitness"
- **Prior density**: $p(\mathbf{w})$
  - Anything we know about parameters before we see any data
- **Marginal likelihood**: $p(\mathbf{y}|\mathbf{X})$
  - It is a normalization constant – ensures $\int p(\mathbf{w}|\mathbf{X}, \mathbf{y}) \, d\mathbf{w} = 1$.

When can we compute the posterior?

Conjugacy (definition)

A prior $p(\mathbf{w})$ is said to be conjugate to a likelihood it results in a posterior of the same type of density as the prior.

- Example:
  - Prior: Gaussian; Likelihood: Gaussian; Posterior: Gaussian
  - Prior: Beta; Likelihood: Binomial; Posterior: Beta
  - Many others...

Why is this important?

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- If prior and likelihood are conjugate, we **know** the form of $p(\mathbf{w}|\mathbf{X},\mathbf{y})$
- Therefore, we **know** the form of the normalizing constant
- Therefore, we **don't need** to compute $p(\mathbf{y}|\mathbf{X})$
- We just need to use some algebra to make $p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})$ **look like** the correct density, ignoring all terms without $\mathbf{w}$

Bayesian Linear Regression - Finding posterior parameters

- Back to our model...
- The posterior must be Gaussian
- Ignoring normalizing constants, the posterior is:

$$\begin{aligned}
p(\mathbf{w}|\mathbf{X},\mathbf{y},\sigma^2) &\propto \exp\left\{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right\} \\
&= \exp\left\{-\frac{1}{2}(\mathbf{w}^\top\boldsymbol{\Sigma}^{-1}\mathbf{w} - 2\mathbf{w}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{w}^\top\boldsymbol{\Sigma}^{-1}\mathbf{w} - 2\mathbf{w}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\}
\end{aligned}$$

Bayesian Linear Regression - Finding posterior parameters

- Ignoring non-$\mathbf{w}$ terms, the prior multiplied by the likelihood is:

$$\begin{aligned}
&p(\mathbf{y}|\mathbf{w},\mathbf{X},\sigma^2) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\right\}\exp\left\{-\frac{1}{2}\mathbf{w}^\top\mathbf{S}^{-1}\mathbf{w}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\mathbf{w}^\top\left[\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}+\mathbf{S}^{-1}\right]\mathbf{w} - \frac{2}{\sigma^2}\mathbf{w}^\top\mathbf{X}^\top\mathbf{y}\right)\right\}
\end{aligned}$$

- Posterior (from previous slide):

$$\propto \exp\left\{-\frac{1}{2}(\mathbf{w}^\top\boldsymbol{\Sigma}^{-1}\mathbf{w} - 2\mathbf{w}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\}$$

Bayesian Linear Regression - Finding posterior parameters

- Equate individual terms on each side.
- Covariance:

$$\mathbf{w}^\top \boldsymbol{\Sigma}^{-1}\mathbf{w} = \mathbf{w}^\top \left[\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \mathbf{S}^{-1}\right]\mathbf{w}$$

$$\boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \mathbf{S}^{-1}\right)^{-1}$$

- Mean:

$$2\mathbf{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \frac{2}{\sigma^2}\mathbf{w}^\top\mathbf{X}^\top\mathbf{y}$$

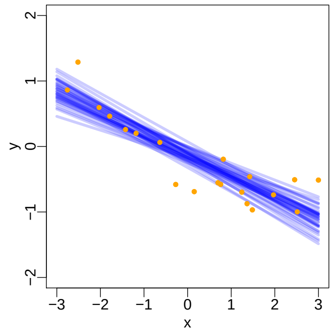
$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\mathbf{X}^\top\mathbf{y}$$

Bayesian Linear Regression - Example

- Linear model with two parameters

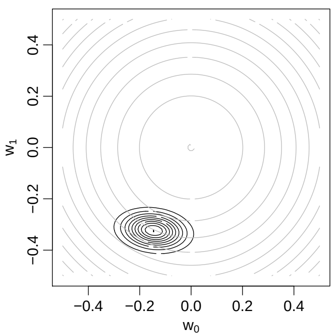
$$f(x) = w_0 + w_1 x$$

- Predictions obtained when sampling from the posterior over parameters



Bayesian Linear Regression - Example

- Posterior distribution over model parameters
- Intercept w_0 and slope w_1



Predictive Distribution

- We can analyze the predictive distribution
- The posterior is central in this analysis

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- as it makes it possible to obtain:

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \int p(y_*|\mathbf{x}_*, \mathbf{w}, \sigma^2)p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2)d\mathbf{w}$$

- Same tedious exercise as before yields:

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \mathcal{N}(y_*|\mathbf{x}_*^\top\boldsymbol{\mu}, \sigma^2 + \mathbf{x}_*^\top\boldsymbol{\Sigma}\mathbf{x}_*)$$

Introducing basis functions

- Imagine transforming the inputs using a set of D functions

$$\mathbf{x} \to \boldsymbol{\varphi}(\mathbf{x}) = (\varphi_1(\mathbf{x}), \ldots, \varphi_D(\mathbf{x}))^\top$$

- The functions $\varphi_1(\mathbf{x})$ are also known as basis functions
- Define:

$$\boldsymbol{\Phi} = \begin{bmatrix} \varphi_1(\mathbf{x}_1) & \ldots & \varphi_D(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(\mathbf{x}_N) & \ldots & \varphi_D(\mathbf{x}_N) \end{bmatrix}$$

Introducing basis functions

- Applying Bayesian Linear Regression on the transformed features gives

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Covariance:

$$\boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \mathbf{S}^{-1}\right)^{-1}$$

- Mean:

$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top\mathbf{y}$$

- Predictions:

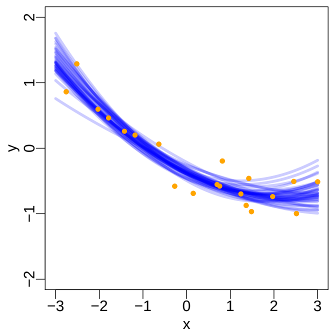
$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \mathcal{N}(y_*|\boldsymbol{\varphi}(\mathbf{x}_*)^\top\boldsymbol{\mu}, \sigma^2 + \boldsymbol{\varphi}(\mathbf{x}_*)^\top\boldsymbol{\Sigma}\boldsymbol{\varphi}(\mathbf{x}_*))$$

Predictions

- Predictions obtained with a polynomial

$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 2$



Computing posterior: recipe

- (Assuming prior conjugate to likelihood)
- Write down prior times likelihood (ignoring any constant terms)
- Write down posterior (ignoring any constant terms)
- Re-arrange them so the look like one another
- Equate terms on both sides to read off parameter values.

Marginal likelihood

- So far, we've ignored $p(\mathbf{y}|\mathbf{X},\sigma^2)$, the normalizing constant in Bayes rule.
- We stated that it was equal to:

$$p(\mathbf{y}|\mathbf{X},\sigma^2) = \int p(\mathbf{y}|\mathbf{X},\mathbf{w},\sigma^2)p(\mathbf{w})\ d\mathbf{w}$$

- We're averaging over all values of $\mathbf{w}$ to get a value for **how good the model is**.
  - How likely is $\mathbf{y}$ given $\mathbf{X}$ and the model
- We can use this to compare models and to optimize σ^2!

Marginal likelihood

- When prior is $\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ and likelihood is $\mathcal{N}(\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})$, marginal likelihood is:

$$p(\mathbf{y}|\mathbf{X},\mathbf{y},\sigma^2,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mu_0,\sigma^2\mathbf{I}+\mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top)$$

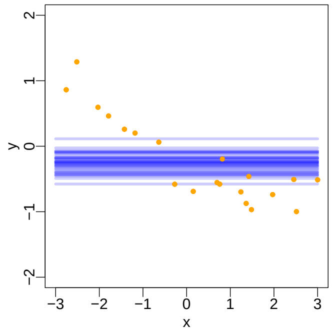
- i.e. an N-dimensional Gaussian evaluated at $\mathbf{y}$.

Model Selection using Marginal Likelihood

- Consider polynomial functions:

$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 0$

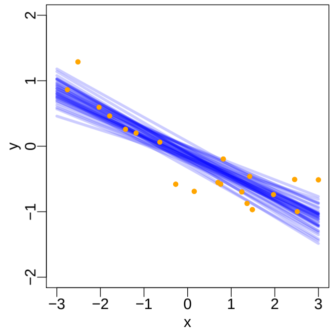


Model Selection using Marginal Likelihood

- Consider polynomial functions:

$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 1$

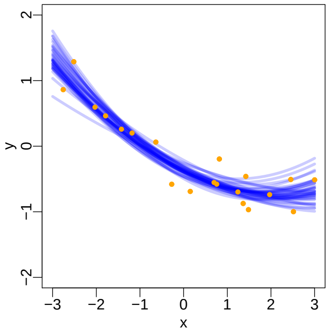


Model Selection using Marginal Likelihood

- Consider polynomial functions:

$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 2$



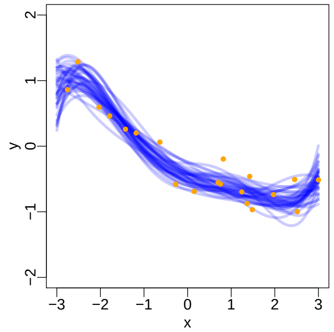
Bayesian Linear Regression
Marginal likelihood

Model Selection using Marginal Likelihood

- Consider polynomial functions:

$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 5$



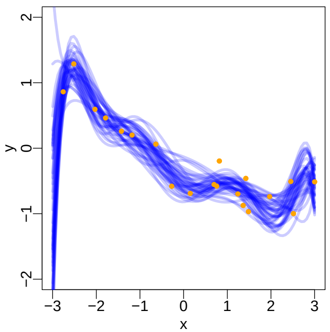
Bayesian Linear Regression
Marginal likelihood

Model Selection using Marginal Likelihood

- Consider polynomial functions:

$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 8$



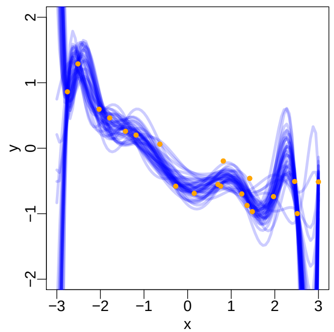
Bayesian Linear Regression
Marginal likelihood

Model Selection using Marginal Likelihood

- Consider polynomial functions:

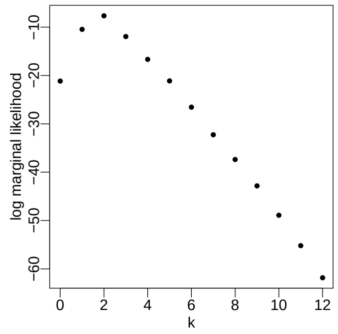
$$f(x) = \sum_{i=0}^{k} w_i x^i$$

- Polynomial with $k = 12$



Model Selection using Marginal Likelihood

- Marginal likelihood as a way to choose the "best" model



Choosing a prior

- How should we choose the prior?
  - Prior effect will diminish as more data arrive.
  - When we don't have much data, prior is very important.
- Some influencing factors:
  - Data type: real, integer, string, etc.
  - Expert knowledge: 'the coin is fair', 'the model should be simple'
  - Computational considerations (not as important as it used to be!)
  - If we know nothing, can use a broad prior – e.g. uniform density.

Summary

- Moved away from a single parameter value.
- Saw how predictions could be made by averaging over all possible parameter values – Bayesian.
- Saw how Bayes rule allows us to get a density for **w** conditioned on the data (and other stuff).
- Computing the posterior is hard except in some cases....
- ....we can do it when things are conjugate.
- Can also (sometimes) compute the marginal likelihood....
- ...and use it for comparing models.
  - No need for costly cross-validation.

Class exercise

- Data: outcomes of N coin tosses (summarized as number of heads) – y_N
- Want a posterior density over r, the probability that a coin toss results in a head.
- Likelihood – binomial:

$$p(y_n|r) = \binom{N}{y_N} r^{y_N}(1-r)^{N-y_N}$$

- Prior – beta:

$$p(r|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}$$

- Beta is **conjugate** to binomial. Therefore posterior is beta. In general, beta is:

$$p(a|c,d) = \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} a^{c-1}(1-a)^{d-1}$$

- Find posterior: $p(r|y_N,\alpha,\beta)$

Solution

- Posterior is proportional to:

$$p(r|y_N, \alpha, \beta) \propto r^{\gamma-1}(1-r)^{\delta-1}$$

- Prior times likelihood is proportional to:

$$\begin{aligned} &\propto \quad r^{\alpha-1}(1-r)^{\beta-1}r^{y_N}(1-r)^{N-y_N} \\ &= \quad r^{y_N+\alpha-1}(1-r)^{N-y_N+\beta-1} \end{aligned}$$

- So:

$$\gamma = y_N + \alpha, \;\; \delta = \beta + N - y_N$$

Class exercise continued...

- By averaging over this posterior over r, we'd like to know the probability of y_* heads in N throws:

$$P(y_*|y_N, \alpha, \beta)$$

- This is an expectation:

$$\begin{aligned} p(y_*|y_N, \alpha, \beta) &= \mathrm{E}_{p(r|y_N,\alpha,\beta)}\left[p(y_*|r)\right] \\ &= \int_0^1 p(y_*|r)p(r|y_N, \alpha, \beta)\; dr \end{aligned}$$

- Where:

$$p(y_*|r) = \left(\begin{array}{c} N \\ y_* \end{array}\right) r^{y_*}(1-r)^{N-y_*}$$

- Can we compute the expectation?

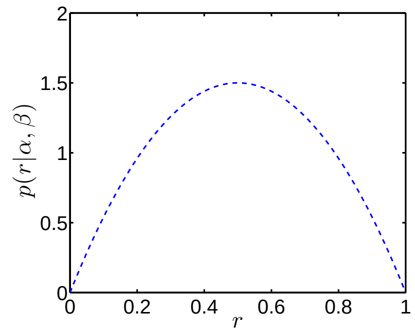
Class exercise continued...

- We don't know what form this will take so cannot ignore constants.

$$\begin{aligned} & p(y_*|y_N, \alpha, \beta) \\ = & \left(\begin{array}{c} N \\ y_* \end{array}\right) \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)} \int_0^1 r^{y_*}(1-r)^{N-y_*}r^{\gamma-1}(1-r)^{\delta-1}\; dr \\ = & \left(\begin{array}{c} N \\ y_* \end{array}\right) \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)} \int_0^1 r^{\gamma+y_*-1}(1-r)^{\delta+N-y_*-1}\; dr \\ = & \left(\begin{array}{c} N \\ y_* \end{array}\right) \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)} \frac{\Gamma(\gamma+y_*)\Gamma(\delta+N-y_*)}{\Gamma(\gamma+y_*+\delta+N-y_*)} \end{aligned}$$

- Where we noticed that the thing in the integral was an unnormalized beta and so its integral must be the inverse of the normalizing constant.

Class exercise – example prior



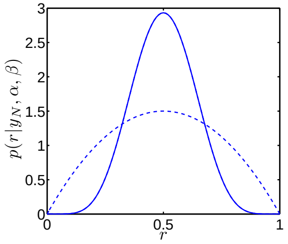
$$\alpha = 2, \;\; \beta = 2$$

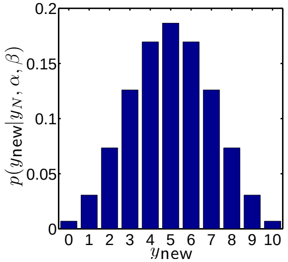
$$p(r|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}$$

$r = 0.5$ is most likely, but we're not sure.

Class exercise – example data

After observing $y_N = 5$ heads in $N = 10$ tosses:



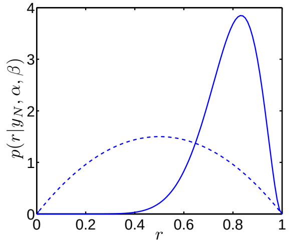


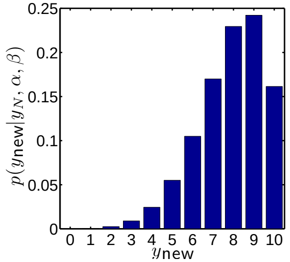
Posterior (left – prior is dashed line) and predictive distribution (right).

Bayesian Linear Regression
Summary

Class exercise – example data 2

After observing $y_N = 9$ heads in $N = 10$ tosses:





Posterior (left – prior is dashed line) and predictive distribution (right).