Bayesian Statistics and Machine Learning

A Conceptual Review

Nicolaj Nørgaard Mühlbach, *Ph.D. student*

Department of Economics and Business Economics, Aarhus University

November 28, 2017

Overview of methodology and purpose

- We will cover the link between Bayesian statistics and machine learning

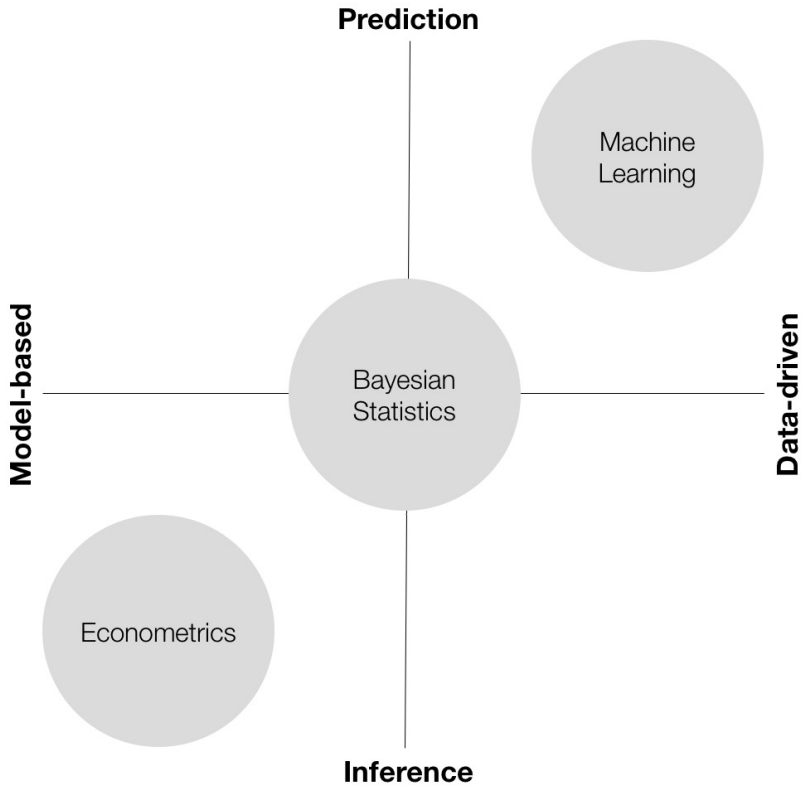
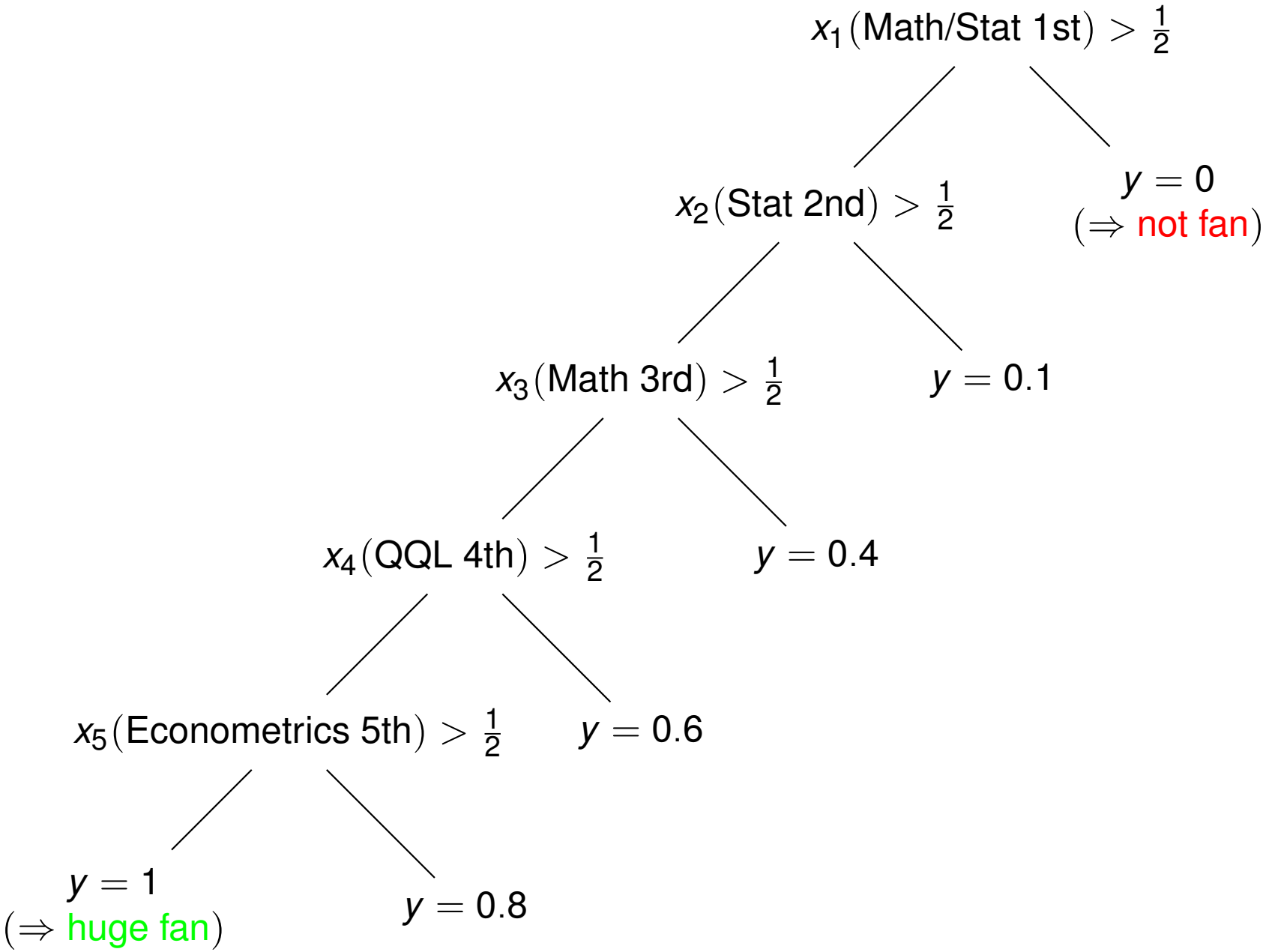


Figure: Different methods and different purposes

How likely are you to be fan of Bayesian/ML methods?

- Assume data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ are samples of B.Sc. Oecon students where:
  - $\mathbf{x}_i \in [0,1]^p$ denotes to which degree student i is fan of the p courses
    - Assume for simplicity that $\frac{1}{N}\sum_{i=1}^N x_{ij} = \frac{1}{2}\ \forall j \in \{1,\dots,p\}$
  - $y_i \in [0,1]$ denotes to which degree student i is fan of the Bayesian ML
- Let $\hat{f}_{\mathcal{D}} : [0,1]^p \mapsto [0,1]$ be *learned* from $\mathcal{D}$ (estimated if you will)
- Now, consider an unseen student $\mathbf{x}_0 \in \mathbb{R}^p$ who could be anyone of you

Question: **Is $\mathbf{x}_0$ likely to be fan of Bayesian ML?**


x_1 (Math/Stat 1st) $> \frac{1}{2}$
x_2 (Stat 2nd) $> \frac{1}{2}$
$y = 0$
($\Rightarrow$ not fan)
x_3 (Math 3rd) $> \frac{1}{2}$
$y = 0.1$
x_4 (QQL 4th) $> \frac{1}{2}$
$y = 0.4$
x_5 (Econometrics 5th) $> \frac{1}{2}$
$y = 0.6$
$y = 1$
($\Rightarrow$ huge fan)
$y = 0.8$

Our goals for this evening

Understand the fundamentals of Bayesian statistics

- Wrap up on fundamental Bayesian theory
- The Bayesian approach to machine learning (or anything)
- The exponential-gamma Bayesian model
- The computational challenge
- Distinctive features of the Bayesian approach
- How Bayesian methods differ from machine learning

Understand the opportunities of machine learning in economics

- Econometric challenges in terms of prediction
- Bias-variance trade-off
- What machine learning does differently

Understand an example of shrinkage models

- The Ridge estimator
- Ridge vs. Lasso from a graphical perspective

Understand The Fundamentals of Bayesian Statistics

"Indeed, if you accept the argument that the false positive rate should be higher for theories that are unlikely, then you have already adopted a fundamentally Bayesian line of reasoning."

Harvey (2017)

Wrap up on fundamental Bayesian theory

- Unit of statistical inference for both frequentists and Bayesians is a family of probability densities

$$\mathcal{F} = \{f_\theta(x); x \in \mathcal{X}, \theta \in \Theta\}, \quad \mathcal{X} \text{ input space, } \Theta \text{ parameter space}$$

- Posterior combines the prior and the conditional likelihood for the parameters
- This is done using Bayes' Rule

$$\begin{aligned} \mathbb{P}(\text{parameters}|\text{data}) &= \frac{\mathbb{P}(\text{parameters}) \times \mathbb{P}(\text{data}|\text{parameters})}{\mathbb{P}(\text{data})} \\ \text{Posterior} &\propto \text{Prior} \times \text{Likelihood} \end{aligned}$$

  - data is fixed at its observed value while θ vary over Θ (opposite of frequentist)

- We make predictions by integrating with respect to the posterior

$$\mathbb{P}(\text{new data}|\text{data}) = \int_{\text{parameters}} \mathbb{P}(\text{new data}|\text{parameters}) \times \mathbb{P}(\text{parameters}|\text{data})$$

- This illustrates how uncertainty is generated in the model

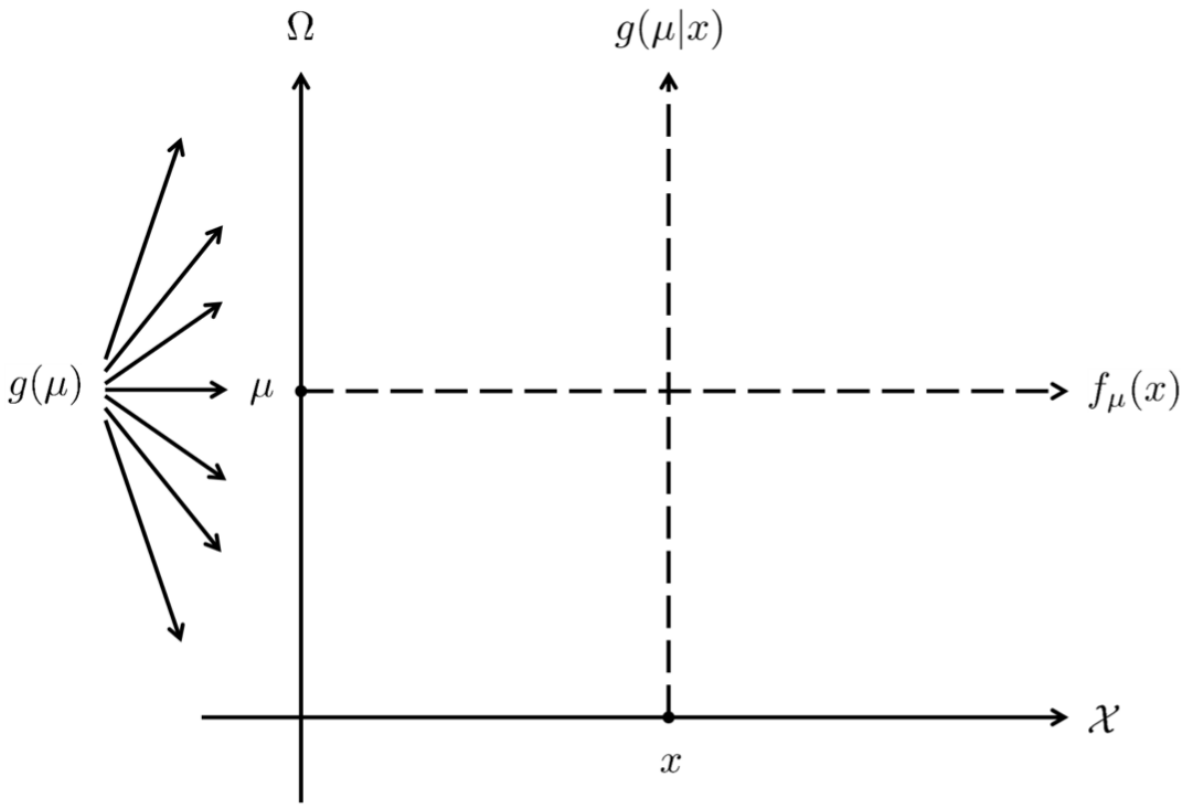


Figure: Bayesian inference proceeds vertically given x; frequentist inference proceeds horizontally given μ

The Bayesian approach to machine learning (or anything)

- *Bayesian modeling* applies Bayes' rule to the unknown variables in a model
  - Many machine learning algorithms involve exactly this step
- In a simple, generic form we can write this process as

$$\begin{array}{lcll} x & \sim & p(x|\theta) & \leftarrow \text{ the data-generating distribution. This is called the } \textit{model} \\ \theta & \sim & p(\theta) & \leftarrow \text{ the a priori knowledge of } \theta\text{. This is called the model } \textit{prior} \end{array}$$

- We want to learn θ from data using

$$p(\theta|x) \underbrace{=}_{\text{exact equal to}} \frac{p(x|\theta)\,p(\theta)}{p(x)} \underbrace{\propto}_{\text{proportional to}} p(x|\theta)\,p(\theta)$$

  - but, unfortunately, $p(\theta|x)$ is unknown.
- However, we have *defined* $p(x|\theta)\,p(\theta)$ and Bayes guides us, although non-trivial:
  - the model $p(x|\theta)$ can be quite complicated
  - $p(\theta|x)$ can be intractable due to the integral in the normalizing constant

The exponential-gamma Bayesian model

- Suppose $X|\lambda \sim \mathcal{E}(\lambda)$, i.e. data is exponentially distributed with parameter λ. Then,

$$f_X(x) = \lambda e^{-\lambda x}, \quad x > 0, \lambda > 0 \quad \text{(Probability density function)}$$

- Exponential distribution describes the time between events in a Poisson process

- We want to learn $f(\lambda|x)$ because $\mathbb{E}[X] = \frac{1}{\lambda}$ and $\mathbb{V}[X] = \frac{1}{\lambda^2}$.

- Assume $\lambda \sim \mathcal{G}(\alpha, \beta)$, i.e. that the constant average rate is gamma-distributed with $\theta = \{\alpha, \beta\}$.

- Then,

$$p(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0 \quad \text{(Probability density function)}$$

Posterior in the exponential-gamma model

- Given sample data $\mathbf{x} \in \mathbb{R}^N$, the posterior distribution is then

$$
\begin{aligned}
p(\lambda|\mathbf{x}) &\propto p(\lambda) \times \ell(\mathbf{x}|\lambda) \\
&= \underbrace{\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}}_{\text{prior (gamma PDF)}} \times \underbrace{\prod_{i=1}^{N} \lambda e^{-\lambda x_i}}_{\text{likelihood}} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \times \lambda^N \left(e^{-\lambda x_1} \times e^{-\lambda x_2} \times \cdots \times e^{-\lambda x_N} \right) \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha+N-1} e^{-\beta\lambda} \times e^{-\lambda(x_1+x_2+\ldots+x_N)} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha+N-1} e^{-\beta\lambda} \times e^{-\lambda N\overline{x}} \\
&\propto \lambda^{\alpha+N-1} e^{-(\beta+N\overline{x})\lambda} \underbrace{\sim}_{\text{distributed as}} \mathcal{G}(\alpha+N, \beta+N\overline{x})
\end{aligned}
$$

- Only sufficient statistics required to transit from prior to posterior
- From this, we may calculate the predictive distribution of a new sample x_{new}

Conjugate priors in general

Definition 1

Conjugate priors

If $\mathcal{F}$ is a class of sampling distributions $f(x|\theta)$ and $\mathcal{P}$ is a class of prior distributions for θ, $p(\theta)$ we say that $\mathcal{P}$ is conjugate to $\mathcal{F}$ if

$$p(\theta|x) \in \mathcal{P} \forall f(\cdot|\theta) \in \mathcal{F} \wedge p(\cdot) \in \mathcal{P}$$

Example

Let $\mathcal{F} = \{\mathcal{E}\}$, i.e. the *family* we draw x from is only the exponential distribution
Let $\mathcal{P} = \{\mathcal{G}\}$, i.e. the *family* we draw λ from is only the gamma distribution
As derived above, $p(\theta|x) \in \mathcal{P}$, i.e. the posterior is also gamma-distributed
Then, the gamma distribution is a conjugate prior to the exponential distribution

Likelihood	Conjugate Prior
Binomial, Negative binomial, Geometric	Beta
Poisson	Gamma

Table: Discrete distributions

The computational challenge

- Except specifying priors, the big challenge is computing the posterior distribution
- There are four main approaches (but we will not cover *variational approximation*):
- **Analytical integration**
  - Use conjugate priors, which combine nicely with the likelihood
  - Usually too much to hope for
- **Laplace approximation**
  - Treat posterior as being approximately Gaussian by choosing $\{\mu, \Sigma\}$ such that

$$p(\theta|X) \approx^{d} \mathcal{N}(\mu, \Sigma)$$

  - Works well when there's a lot of data and low model complexity
- **Monte Carlo integration (Gibbs sampling or MCMC)**
  - Simulating a Markov chain that eventually converges to the posterior distribution. Imagine we are interested in $\mathbb{E}_{p(X)}[f(X)]$. The idea follows

$$\text{Sample } X_1, \ldots, X_N \sim^{\mathcal{IID}} p(X) \Longrightarrow \mathbb{E}_{p(X)}[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} f(X_i)$$

  - As $N \to \infty$, the approximation converges to the true expectation
  - Can be applied to a remarkable variety of problems (dominant approach)

Distinctive features of the Bayesian approach

Probability

- Probability is used not only to describe "physical" randomness, but also to describe uncertainty regarding the *true* values of the parameters
- These prior and posterior probabilities represent degrees of belief

Modeling

- The Bayesian approach takes modeling seriously
- A Bayesian model includes a suitable prior distribution for model parameters
  - If the model/prior are chosen without regard for the actual situation, there is no justification for believing the results of Bayesian inference

Finite Sample Justification

- The model and prior are chosen based on our knowledge of the problem
- Thus, we do not rely on asymptotic theory, and we do not restrict the complexity of the model just because we have only a small amount of data

Also, features include **model complexity justification** and **problem breakdown**

Model complexity justification

- Image a complex and high-dimensional model
- Often, the log-likelihood will be flat along some dimension
- Adding prior knowledge corresponds to adding curvature to the likelihood

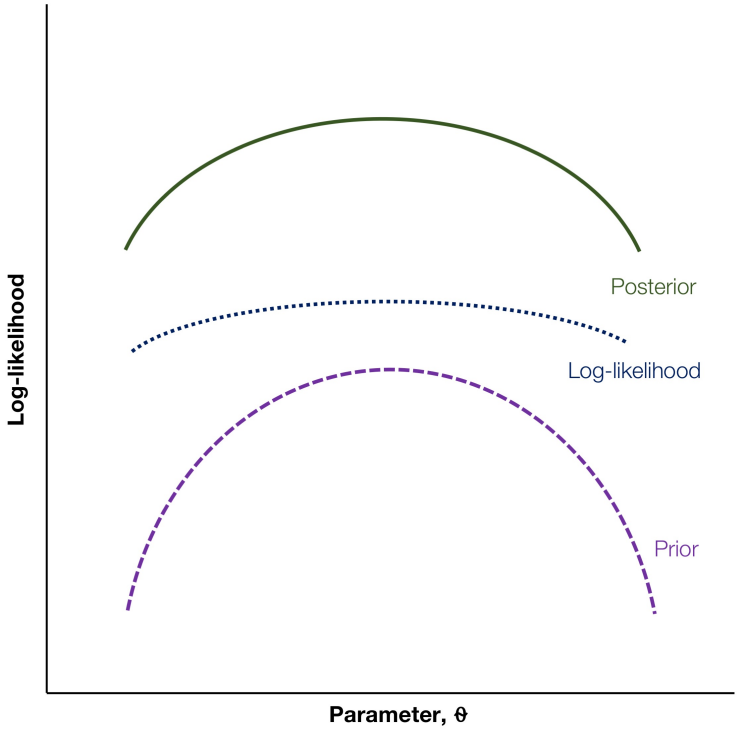


Figure: An Illustration of Bayesian Estimation

Problem breakdown

- Imagine a management consultant aiming at increasing the profits for a client
- Image a researcher trying to estimate the marginal propensity to consume
- How would you proceed?

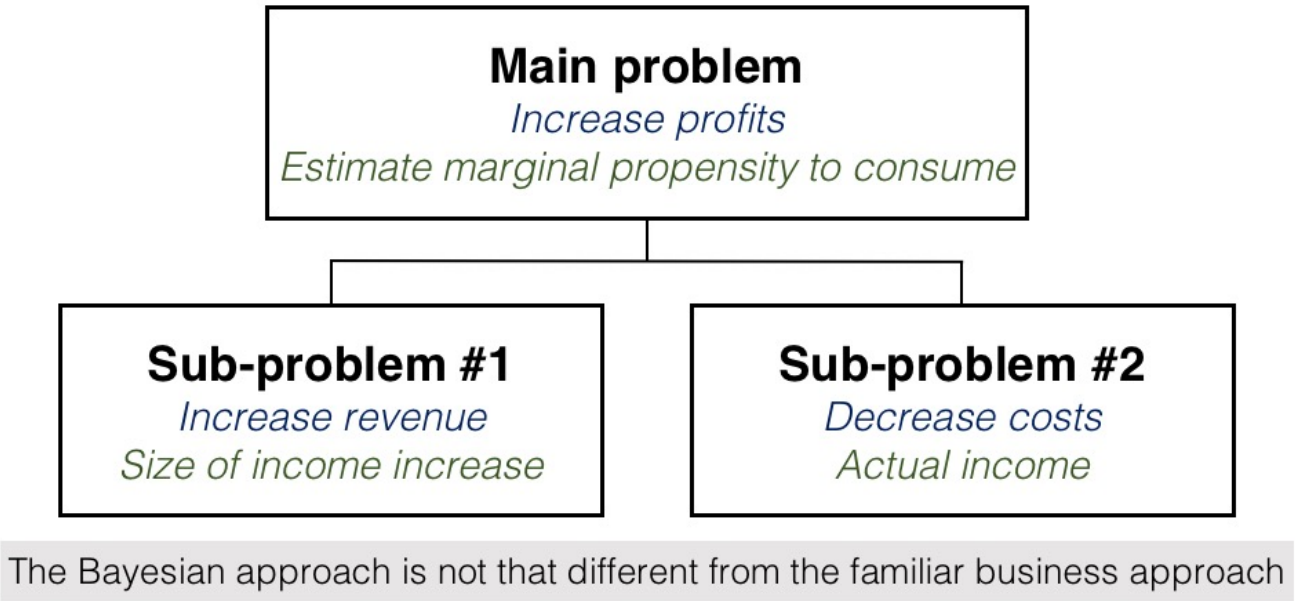


Figure: The Bayesian approach from a business perspective

How Bayesian methods differ from machine learning

- Pure machine learning arises from statistical learning, best illustrated as follows

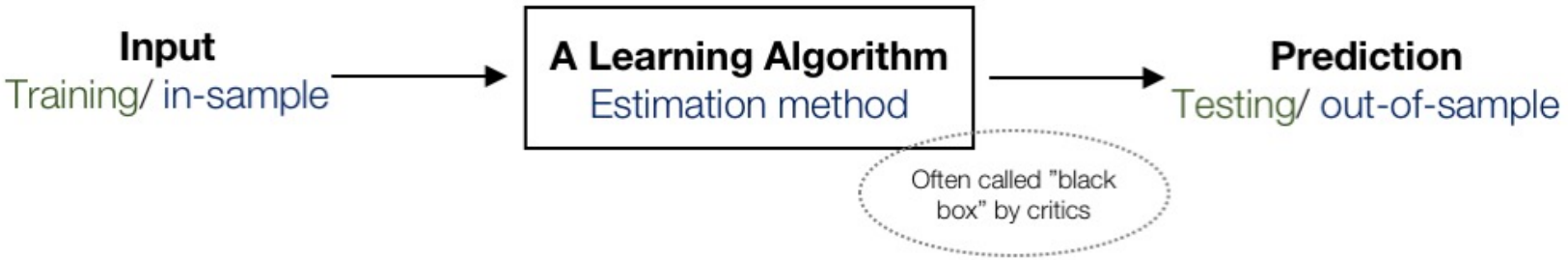


Figure: An Illustration of Machine Learning

- The learning machine has various knobs, whose settings change the prediction
  - Learning is about twiddling the knobs to make better predictions
- This differs profoundly from the Bayesian view, as the learning is arbitrary
- Unlike a model, the machine has no meaningful semantics compared to beliefs
- The knobs do not correspond to the parameters of a Bayesian model

Opportunities of Machine Learning in Economics

"There have been very fruitful collaborations between computer scientists and statisticians in the last decade or so, and I expect collaborations between computer scientists and econometricians will also be productive in the future."

Varian (2014)

Focus of econometricians is to draw inference ...

- Econometricians rely on statistical properties to draw inference
- Suppose model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ and data $\mathbf{X} \in \mathbb{R}^{N \times p}$ with $\mathbb{P}_\theta(\mathbf{X}) = \mathbb{P}(\mathbf{X}|\theta)$
- An estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$ is an **unbiased** estimator for the parameter θ iff

$$\mathbb{E}_\theta\left[\hat{\theta}\right] = \theta \quad \forall \theta \in \Theta$$

- The sequence of estimators $\hat{\theta}_N = \hat{\theta}_N(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is a **consistent** iff

$$\forall \varepsilon > 0 : \lim_{N \to \infty} \mathbb{P}_\theta\left(\left|\hat{\theta}_N - \theta\right| < \varepsilon\right) = 1 \quad \forall \theta \in \Theta$$

- An efficient estimator is **efficient** among the unbiased estimators if

$$\mathbb{V}_\theta\left(\hat{\theta}\right) = \mathcal{I}_\theta^{-1} \quad \forall \theta \in \Theta \quad \text{(Cramér-Rao bound)}$$

... whereas the focus in machine learning is prediction

- Let $\mathcal{L}(f) = \mathbb{E}_{\mathbb{P}(X,Y)}\left[\ell\left(f\left(X\right), Y\right)\right]$, and let $\left\{f^*, \hat{f}\right\}$ denote optimal and feasible
- The objective in machine learning is to minimize the error of prediction given by

$$\underbrace{\mathcal{L}\left(\hat{f}_{\mathcal{D}}\right)}_{\text{prediction error}} = \underbrace{\mathcal{L}\left(\hat{f}_{\mathcal{D}}\right) - \mathcal{L}\left(f^*\right)}_{\text{estimation error}} + \underbrace{\mathcal{L}\left(f^*\right) - \mathcal{L}^*}_{\text{approximation error}} + \underbrace{\mathcal{L}^*}_{\text{irreducible error}}$$

- In causal inference, the objective is to minimize the in-sample prediction error

$$\underbrace{\mathcal{L}\left(\hat{f}_{\mathcal{D}}\right) - \mathcal{L}\left(f^*\right)}_{\text{estimation error}} = \underbrace{\hat{\mathcal{L}}\left(\hat{f}_{\mathcal{D}}\right) - \hat{\mathcal{L}}\left(f^*\right)}_{\text{in-sample prediction error}} + \underbrace{\mathcal{L}\left(\hat{f}_{\mathcal{D}}\right) - \hat{\mathcal{L}}\left(\hat{f}_{\mathcal{D}}\right)}_{\text{unseen overfit}} + \underbrace{\hat{\mathcal{L}}\left(f^*\right) - \mathcal{L}\left(f^*\right)}_{\text{random variation}}$$

- Coming at the cost of unseen overfit and often increases in approximation error

Common specification reveals bias-variance trade-off

- Consider input and output space $\mathcal{X}, \mathcal{Y}$, assume $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ stochastic variables distributed according to unknown joint probability distribution $\mathbb{P}(X, Y)$,
- Specify loss function $\left(\ell(z) = z^2\right)$, common to econometricians
- Observe data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$, $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ $\forall i = \{1, \ldots, N\}$, $\mathcal{D} \sim^{iid} \mathbb{P}(X, Y)$
- Given $\mathcal{D}$, we estimate a function $\hat{f} \in \mathcal{F}$, $\hat{f} : \mathcal{X} \mapsto \mathcal{Y}$ that predicts Y from X.

$$\mathcal{L}\left(\hat{f}\right) = \mathbb{E}_{\mathbb{P}(X,Y)}\left[\left(Y - \underbrace{f^*(X) + f^*(X)}_{\text{artifically added}} - \hat{f}(X)\right)^2\right]$$

$$\underbrace{\mathbb{E}_{\mathbb{P}(X,Y)}\left[\left(Y - f^*(X)\right)^2\right]}_{\text{irreducible noise}} + \underbrace{\mathbb{E}_{\mathbb{P}(X,Y)}\left[\left(f^*(X) - \hat{f}(X)\right)^2\right]}_{\text{mean squared error}}$$

$$+2\underbrace{\mathbb{E}_{\mathbb{P}(X,Y)}\left[\left(Y - f^*(X)\right)\left(f^*(X) - \hat{f}(X)\right)\right]}_{\text{covariance of noise and bias}}.$$

Irreducible error and covariance of noise and bias

- Denote by σ_ε^2 the irreducible noise $\mathbb{E}_{\mathbb{P}(X,Y)}\left[\left(Y - f^*\left(X\right)\right)^2\right]$
- The covariance of noise and bias vanishes due to

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}(X,Y)}\left[\left(Y - f^*\left(X\right)\right)\left(f^*\left(X\right) - \hat{f}\left(X\right)\right)\right] &= \\
\int\int \left(Y - f^*\left(X\right)\right)\left(f^*\left(X\right) - \hat{f}\left(X\right)\right) \underbrace{\mathbb{P}\left(Y|X\right)\mathbb{P}\left(X\right)}_{=\mathbb{P}(Y,X)} dY dX &= \\
\int \underbrace{\left\{\mathbb{E}_{Y|X}\left(Y - f^*\left(X\right)\right)\right\}}_{=0}\left(f^*\left(X\right) - \hat{f}\left(X\right)\right)\mathbb{P}\left(X\right) dX,&
\end{aligned}$$

Mean squared error and bias-variance trade-off

$$\begin{aligned}
\text{MSE} &= \mathbb{E}_{\mathbb{P}(X,Y)}\left[\left(f^*(X) - \mathbb{E}_{\mathcal{D}}\left[\hat{f}(X)\right] + \mathbb{E}_{\mathcal{D}}\left[\hat{f}(X)\right] - \hat{f}(X)\right)^2\right] \\
&= \mathbb{E}_{\mathbb{P}(X,Y)}\left[\left(f^*(X) - \mathbb{E}_{\mathcal{D}}\left[\hat{f}(X)\right]\right)^2 + \left(\mathbb{E}_{\mathcal{D}}\left[\hat{f}(X)\right] - \hat{f}(X)\right)^2\right. \\
&\quad \left. +2\left(f^*(X) - \mathbb{E}_{\mathcal{D}}\left[\hat{f}(X)\right]\right)\left(\mathbb{E}_{\mathcal{D}}\left[\hat{f}(X)\right] - \hat{f}(X)\right)\right].
\end{aligned}$$

- Expectation with respect to $\mathcal{D}$ (Fubini's Theorem) causes the last term to vanish

$$\begin{aligned}
\mathbb{E}_{\mathcal{D},\mathbb{P}(X,Y)}\left[\left(Y - \hat{f}(X)\right)^2\right] &= \\
\sigma_\varepsilon^2 & \quad \text{(noise variance)} \\
+\mathbb{E}_{\mathcal{D},\mathbb{P}(X,Y)}\left[\left(f^*(X) - \mathbb{E}_{\mathcal{D}}\left[\hat{f}(X)\right]\right)^2\right] & \quad \text{(expected squared bias)} \\
+\mathbb{E}_{\mathcal{D},\mathbb{P}(X,Y)}\left[\left(\mathbb{E}_{\mathcal{D}}\left[\hat{f}(X)\right] - \hat{f}(X)\right)^2\right] & \quad \text{(expected variance)}.
\end{aligned}$$

The bias-variance trade-off illustrates the different objectives

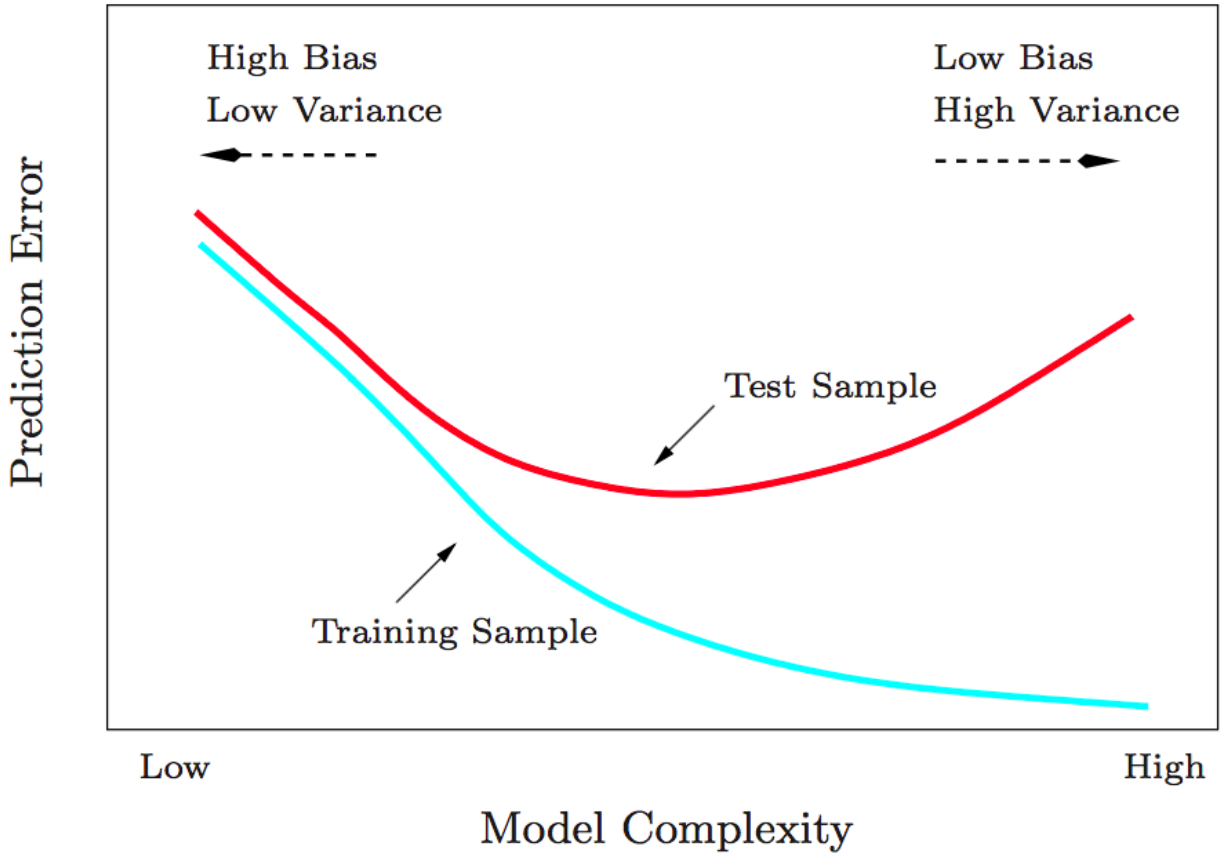


Figure: Bias-variance trade-off

What does machine learning differently?

- Function class $\mathcal{F}$ much wider, $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathcal{Y}\}$ compared to e.g. OLS
- Regularization to avoid overfit, $\mathcal{A} : (\lambda, \Theta, \mathcal{D}) \mapsto \hat{f}_{\mathcal{D}} \in \mathcal{F}$

$$\hat{f}_{\mathcal{D}} = \arg\min_{f \in \mathcal{F}} \hat{\mathcal{L}}(f) + \lambda \mathcal{J}(f) \quad \text{(Tikhonov regularization)}$$

- The role of λ, and tuning via cross-validation:

Step 1 Split the training data $\mathcal{D}$ into k equal-sized folds with $\frac{N}{k}$ observations per fold

Step 2 Denote by $\hat{f}_{\mathcal{D}}^{-\kappa(i)}(\mathbf{x}_i)$ the prediction for observation $(y_i, \mathbf{x}_i)$ on other folds

Step 3 Calculate the CV error as the average prediction loss on the left-out fold $j \in \{1, \ldots, k\}$:

$$CV_j(\Theta, \lambda) = \frac{1}{N_k} \sum_{\{i:\kappa(i)=j\}} \ell\left(\hat{f}_{\mathcal{D}}^{-\kappa(i)}\right)$$

Step 4 Iterate steps 2-3 for each of the k folds to obtain a k vector of $CV_j(\Theta, \lambda)$, which is to be averaged as:

$$CV(\Theta, \lambda) = \frac{1}{k} \sum_{j=1}^{k} \left[\frac{1}{N_k} \sum_{\{i:\kappa(i)=j\}} \ell\left(\hat{f}_{\mathcal{D}}^{-\kappa(i)}\right) \right] = \frac{1}{N} \sum_{i=1}^{N} \ell\left(\hat{f}_{\mathcal{D}}^{-\kappa(i)}\right)$$

An example: Shrinkage models

The Ridge estimator

- Observe data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$, $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ $\forall i = \{1, \ldots, N\}$, $\mathcal{D} \sim^{iid} \mathbb{P}(X, Y)$

$$\hat{\mathcal{L}}^{\text{Ridge}}\left(\hat{\boldsymbol{\beta}}, \lambda\right) = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^{\top}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) + \lambda\hat{\boldsymbol{\beta}}^{\top}\hat{\boldsymbol{\beta}} = \mathbf{y}^{\top}\mathbf{y} - 2\hat{\boldsymbol{\beta}}^{\top}\mathbf{X}^{\top}\mathbf{y} + \hat{\boldsymbol{\beta}}^{\top}\mathbf{X}^{\top}\mathbf{X}\hat{\boldsymbol{\beta}} + \lambda\hat{\boldsymbol{\beta}}^{\top}\hat{\boldsymbol{\beta}}$$

- FOC implies that

$$\begin{aligned}\frac{\partial\hat{\mathcal{L}}(\boldsymbol{\beta}, \lambda)}{\partial\hat{\boldsymbol{\beta}}^{\top}} &= -2\mathbf{X}^{\top}\mathbf{y} + 2\mathbf{X}^{\top}\mathbf{X}\hat{\boldsymbol{\beta}} + 2\lambda\hat{\boldsymbol{\beta}} = 0 \Leftrightarrow \left(\mathbf{X}^{\top}\mathbf{X} + \lambda\mathbb{I}_p\right)\hat{\boldsymbol{\beta}} = \mathbf{X}^{\top}\mathbf{y} \\ \hat{\boldsymbol{\beta}}^{\text{ridge}} &= \left(\mathbf{X}^{\top}\mathbf{X} + \lambda\mathbb{I}_p\right)^{-1}\mathbf{X}^{\top}\mathbf{y}\end{aligned}$$

which essentially means that

$$\underbrace{\mathbb{E}\left[\hat{\boldsymbol{\beta}}^{\text{ridge}}|\mathbf{X}\right] \leq \boldsymbol{\beta} = \mathbb{E}\left[\hat{\boldsymbol{\beta}}^{\text{OLS}}|\mathbf{X}\right]}_{\text{bias increases}} \quad \wedge \quad \underbrace{\text{tr}\left(\mathbb{V}\left[\hat{\boldsymbol{\beta}}^{\text{ridge}}|\mathbf{X}\right]\right) \leq \text{tr}\left(\mathbb{V}\left[\hat{\boldsymbol{\beta}}^{\text{OLS}}|\mathbf{X}\right]\right)}_{\text{variance decreases}}$$

- Thus, we **intentionally** introduce a bias, which decreases the variance
  - in addition, $\mathbf{X}^{\top}\mathbf{X} + \lambda\mathbb{I}_p$ is generally non-singular even if $\mathbf{X}^{\top}\mathbf{X}$ is singular

Ridge vs. Lasso from a graphical perspective

- The choice of ℓ_2-norm is somewhat arbitrary. An extension is the LASSO

$$\text{Lasso: } \hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t \quad (\text{using } \ell_1\text{-norm})$$

$$\text{Ridge: } \hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq t \quad (\text{using squared } \ell_2\text{-norm})$$

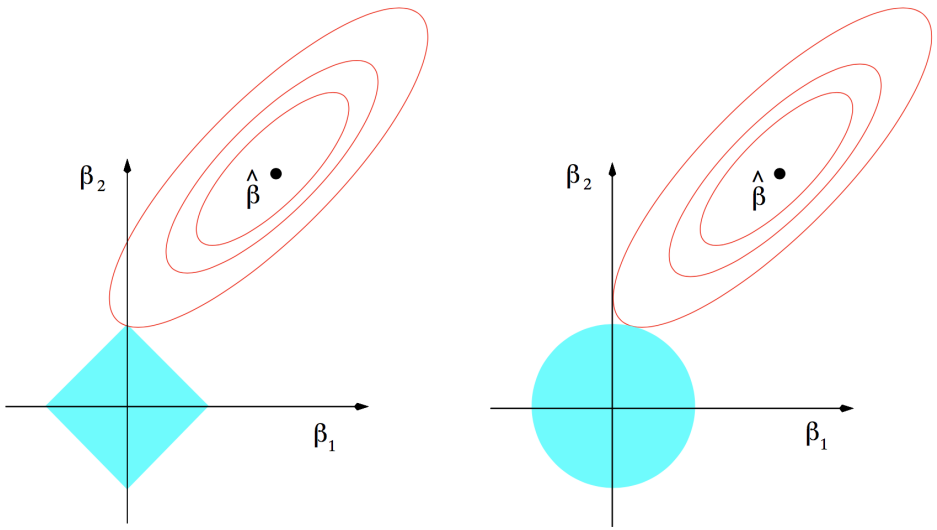


Figure: Lasso (left) and ridge (right) visualizations

Oracle Properties

- Suppose $\mathbb{E}[Y|\mathbf{X}] = \beta_1 X_1 + \ldots + \beta_p X_p$, and that $\mathcal{A} = \{j : \beta_j \neq 0\}, \quad |\mathcal{A}| < p$.

- One can show that the Lasso enjoys the first *Oracle* property (but not the second!)

$$\mathbb{P}\left(\hat{\beta}^{\text{Lasso}}_{\mathcal{A}^c} = 0\right) \to 1$$