Bayesian Prediction and Decision Making
Pizza and beer talk at iZettle

Lecture 4: Predictions. Decisions.

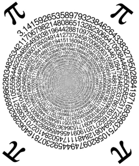
Mattias Villani

Department of Statistics
Stockholm University
and
Department of Computer and Information Science
Linköping University

- **A two slide intro to Bayesian Learning**
- **Bayesian Prediction**
- **Bayesian Decision Making**

Uncertainty and subjective probability

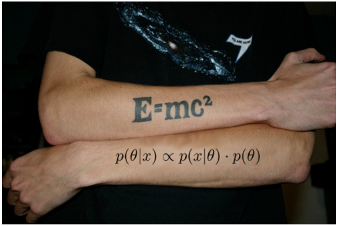
- What we really want: $\Pr(\text{unknown}|\text{known})$
  - $\Pr(\theta|\text{data})$
  - $\Pr(\text{test data}|\text{training data})$
- $\Pr(\theta < 0.6|\text{data})$ only makes sense if θ is random.
- But θ may be a fixed natural constant?
- **Bayesian: doesn't matter if θ is fixed or random**.
- Do **You** know the value of θ or not?
- $p(\theta)$ reflects Your knowledge/**uncertainty** about θ.
- **Subjective probability**.
- The statement $\Pr(\text{10th decimal of } \pi = 9) = 0.1$ makes sense.



- **Bayesian learning** combines:
  - prior information $p(\theta)$ with
  - data information $p(Data|\theta)$ (**likelihood** function)
  - using Bayes theorem

$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{p(Data)} \propto p(Data|\theta)p(\theta)$$

Posterior $\propto$ Likelihood $\cdot$ Prior



- **Posterior predictive density** for future $\tilde{y}$ given observed $\mathbf{y}$

$$p(\tilde{y}|\mathbf{y}) = \int_{\theta} p(\tilde{y}|\theta, \mathbf{y}) p(\theta|\mathbf{y}) d\theta$$

- If $p(\tilde{y}|\theta, \mathbf{y}) = p(\tilde{y}|\theta)$ [not true for time series], then

$$p(\tilde{y}|\mathbf{y}) = \int_{\theta} p(\tilde{y}|\theta) p(\theta|\mathbf{y}) d\theta$$

- **Parameter uncertainty** in $p(\tilde{y}|\mathbf{y})$ by **averaging over** $p(\theta|\mathbf{y})$.

- **Simulation** implementation:
  - Simulate from posterior $\theta^{(i)} \sim p(\theta|\mathbf{y})$
  - Simulate $\tilde{y}^{(i)} \sim p(y|\theta^{(i)})$ from model

- **Autoregressive process**

$$y_t \quad = \quad \mu + \phi_1(y_{t-1} - \mu) + \ldots + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \ \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

Simulation algorithm. Repeat N times:

1. Generate a **posterior draw** of $\theta^{(1)} = (\phi_1^{(1)}, \ldots, \phi_p^{(1)}, \mu^{(1)}, \sigma^{(1)})$ from $p(\phi_1, \ldots, \phi_p, \mu, \sigma | \mathbf{y}_{1:T})$.
2. Generate a **predictive draw** of future time series by:
   2.1 $\tilde{y}_{T+1} \sim p(y_{T+1} | y_T, y_{T-1}, \ldots, y_{T-p}, \theta^{(1)})$
   2.2 $\tilde{y}_{T+2} \sim p(y_{T+2} | \tilde{y}_{T+1}, y_T, \ldots, y_{T-p}, \theta^{(1)})$
   2.3 $\tilde{y}_{T+3} \sim p(y_{T+3} | \tilde{y}_{T+2}, \tilde{y}_{T+1}, y_T, \ldots, y_{T-p}, \theta^{(1)})$
   2.4 ...

Binary classification

- Response is assumed to be **binary** ($y = 0$ or 1).
- Example: Spam/Ham. Covariates: $-symbols, etc.
- **Logistic regression**

$$\Pr(y_i = 1 \mid x_i) = \frac{\exp(\mathbf{x}_i'\beta)}{1 + \exp(\mathbf{x}_\mathbf{i}'\beta)}$$

- **Multi-class** ($c = 1, 2, ..., C$) logistic regression

$$\Pr(y_i = c \mid x_i) = \frac{\exp(\mathbf{x}_\mathbf{i}'\beta_c)}{\sum_{k=1}^{C} \exp(\mathbf{x}_i'\beta_k)}$$

- **Likelihood logistic regression**

$$p(\mathbf{y}|\mathbf{X}, \beta) = \prod_{i=1}^{n} \frac{[\exp(\mathbf{x}_i'\beta)]^{y_i}}{1 + \exp(\mathbf{x}_i'\beta)}.$$

- Posterior is non-standard. What to do?

- **Normal approximation**
  - Use $\theta \overset{approx}{\sim} N(\hat{\theta}, \Omega)$
  - $\hat{\beta}$ is the mode of the posterior
  - $\Omega = -H^{-1}$, where H is the Hessian matrix at the mode

$$\Omega = -\frac{\partial^2 \ln p(\theta|\mathbf{y})}{\partial\theta\partial\theta^T}|_{\theta=\tilde{\theta}}.$$

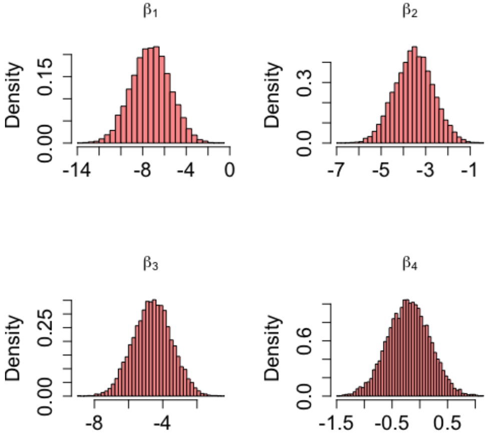
  - Theory: the posterior will be $N(\hat{\theta}, \Omega)$ is large datasets.
  - Both $\hat{\theta}$ and H can be obtained with **numerical optimization**.
  - Only need to code $\log p(\mathbf{y}|\theta) + \log p(\theta)$

- Markov Chain Monte Carlo (**MCMC**) or Hamiltonian MC (**HMC**).

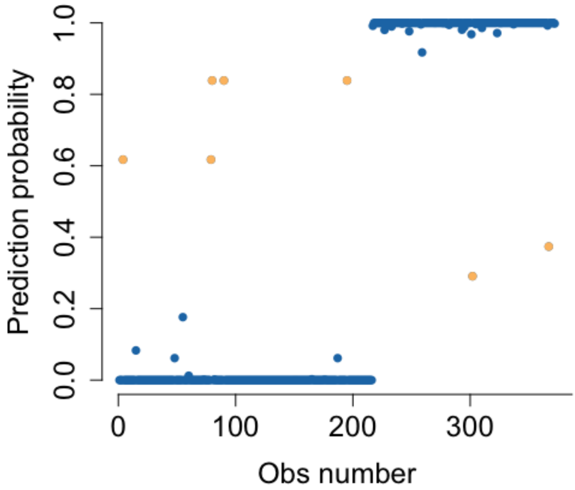
- **Variational inference**: use optimization to find a simpler distribution $q(\theta)$ that minimizes the (Kullback-Leibler) distance between $q(\theta)$ and $p(\theta|\mathbf{y})$.

Posterior - fraud data

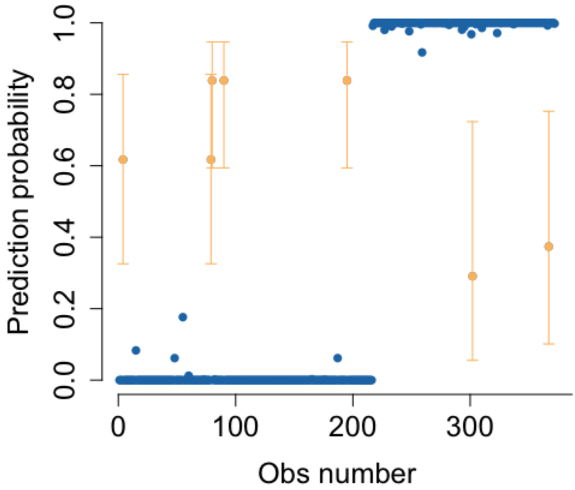
- Predicting fraudulent bills from 4 image features.
- Logistic regression.
- nTrain = 1000, Test = 372.



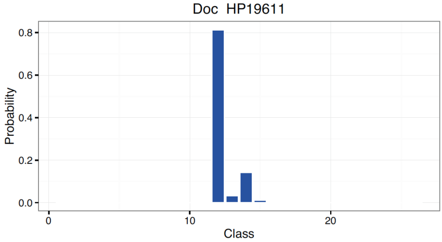
The ask-a-human option - fraud

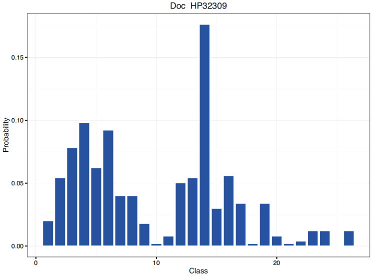


The ask-a-human option - fraud



The ask-a-human option - bug allocation





- Let θ be an **unknown quantity**. **State of nature**. Examples: Future inflation, Global temperature, Fraud.
- Let $a \in \mathcal{A}$ be an **action**. Ex: Interest rate, Energy tax, Surgery.
- Choosing action a when state of nature is θ gives **utility**

$$U(a, \theta)$$

- Alternatively loss $L(a, \theta) = -U(a, \theta)$.
- Loss table:

	θ_1	θ_2
a_1	$L(a_1, \theta_1)$	$L(a_1, \theta_2)$
a_2	$L(a_2, \theta_1)$	$L(a_2, \theta_2)$

- Example:

	Rainy	Sunny
Umbrella	20	10
No umbrella	50	0

- Example:
  - θ is the number of items demanded of a product
  - a is the number of items in stock
  - Utility

$$U(a,\theta) = \begin{cases} p \cdot \theta - c_1(a-\theta) & \text{if } a > \theta \text{ [too much stock]} \\ p \cdot a - c_2(\theta - a)^2 & \text{if } a \leq \theta \text{ [too little stock]} \end{cases}$$

Optimal decision

- Ad hoc decision rules: *Minimax*. *Minimax-regret* etc etc ...
- **Bayesian theory**: maximize the **posterior expected utility**:

$$a_{bayes} = \text{argmax}_{a \in \mathcal{A}}\ E_{p(\theta|y)}[U(a, \theta)],$$

  where $E_{p(\theta|y)}$ denotes the posterior expectation.
- Using simulated draws $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ from $p(\theta|y)$:

$$E_{p(\theta|y)}[U(a, \theta)] \approx N^{-1} \sum_{i=1}^{N} U(a, \theta^{(i)})$$

- **Separation principle**:

1. First obtain $p(\theta|y)$
2. then form $U(a, \theta)$ and finally
3. choose a that maximes $E_{p(\theta|y)}[U(a, \theta)]$.