Machine Learning - MT 2017
7 Bayesian Approach to Machine Learning

Christoph Haase

University of Oxford
October 23, 2017

Frequentist vs Bayesian Approaches

Different views on probability:

- Frequentists: Probability of an event represents long-run frequency over a large number of repetitions of an experiment
- Bayesians: Probability of an event represents a degree of belief about the event

Frequentist vs Bayesian Approaches

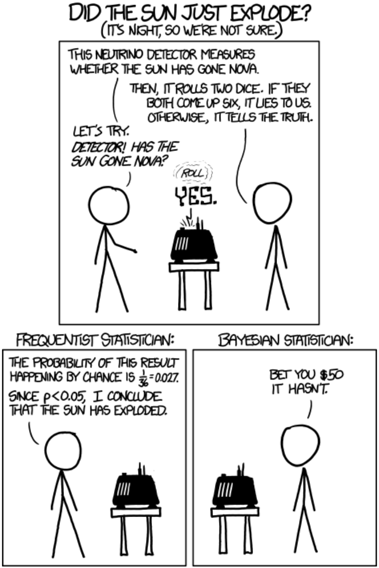
Different views on probability:

- Frequentists: Probability of an event represents long-run frequency over a large number of repetitions of an experiment
- Bayesians: Probability of an event represents a degree of belief about the event

Different views on statistics:

- Frequentists: Parameters are fixed, data are a repeatable random sample, underlying parameters remain constant at every repetition
- Bayesians: Data are fixed, parameters are unknown and described probabilistically, repetition adds knowledge about parameters

Frequentist vs Bayesian Approaches



Bayes' Theorem

Recall basic laws of probability:

$$p(A \cap B)$$

Bayes' Theorem

Recall basic laws of probability:

$$p(A \cap B) = p(A|B) \cdot p(B)$$

Bayes' Theorem

Recall basic laws of probability:

$$p(B|A) \cdot p(A) = p(A \cap B) = p(A|B) \cdot p(B)$$

Bayes' Theorem

Recall basic laws of probability:

$$p(B|A) \cdot p(A) = p(A \cap B) = p(A|B) \cdot p(B)$$

Bayes' Theorem:

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{P(B)}$$

Bayes' Theorem

Recall basic laws of probability:

$$p(B|A) \cdot p(A) = p(A \cap B) = p(A|B) \cdot p(B)$$

Bayes' Theorem:

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{P(B)}$$

Viewing A as a proposition and B as evidence:

- $p(A)$ is the prior representing initial belief about A
- $p(A|B)$ is the posterior representing belief about A after learning about B
- Posterior is proportional to prior times likelihood if we fix B:

$$p(A|B) \propto p(B|A) \cdot p(A)$$

Priors Matter

Suppose we have a test for a disease:

- test is 95% effective, i.e., $p(T|D) = 0.95$
- rate of false positives is 1%, i.e., $p(T|\bar{D}) = 0.01$
- the disease occurs in 0.5% of the population, i.e., $p(D) = 0.005$

Priors Matter

Suppose we have a test for a disease:

- test is 95% effective, i.e., $p(T|D) = 0.95$
- rate of false positives is 1%, i.e., $p(T|\bar{D}) = 0.01$
- the disease occurs in 0.5% of the population, i.e., $p(D) = 0.005$

Suppose the test is positive, what is $p(D|T)$:

Priors Matter

Suppose we have a test for a disease:

- test is 95% effective, i.e., $p(T|D) = 0.95$
- rate of false positives is 1%, i.e., $p(T|\bar{D}) = 0.01$
- the disease occurs in 0.5% of the population, i.e., $p(D) = 0.005$

Suppose the test is positive, what is $p(D|T)$:

$$
\begin{aligned}
p(D|T) &= \frac{p(T|D) \cdot p(D)}{p(T)} \\
&= \frac{p(T|D) \cdot p(D)}{p(T|D) \cdot p(D) + p(T|\bar{D}) \cdot p(\bar{D}))} \\
&= \frac{0.95 \cdot 0.005}{0.95 \cdot 0.005 + 0.01 \cdot 0.995} \\
&\approx 0.32
\end{aligned}
$$

Bayesian Machine Learning

In the discriminative framework, we model the output y as a probability distribution given the input $\mathbf{x}$ and the parameters $\mathbf{w}$, say $p(y \mid \mathbf{w}, \mathbf{x})$

In Bayesian machine learning, we assume a prior on the parameters $\mathbf{w}$, say $p(\mathbf{w})$

This prior represents a "belief" about the model; the uncertainty in our knowledge is expressed mathematically as a probability distribution

Bayesian Machine Learning

In the discriminative framework, we model the output y as a probability distribution given the input $\mathbf{x}$ and the parameters $\mathbf{w}$, say $p(y \mid \mathbf{w}, \mathbf{x})$

In Bayesian machine learning, we assume a prior on the parameters $\mathbf{w}$, say $p(\mathbf{w})$

This prior represents a "belief" about the model; the uncertainty in our knowledge is expressed mathematically as a probability distribution

When observations, $\mathcal{D} = \langle (\mathbf{x}_i, y_i) \rangle_{i=1}^{N}$ are made the belief about the parameters $\mathbf{w}$ is updated using Bayes' rule

As before, the posterior distribution on $\mathbf{w}$ given the data $\mathcal{D}$ is:

$$p(\mathbf{w} \mid \mathcal{D}) \propto p(\mathbf{y} \mid \mathbf{w}, \mathbf{X}) \cdot p(\mathbf{w})$$

Coin Toss Example

Let us consider the Bernoulli model for a coin toss, for $\theta \in [0, 1]$

$$p(\mathsf{H} \mid \theta) = \theta$$

Suppose after three independent coin tosses, you get T, T, T. What is the maximum likelihood estimate for θ?

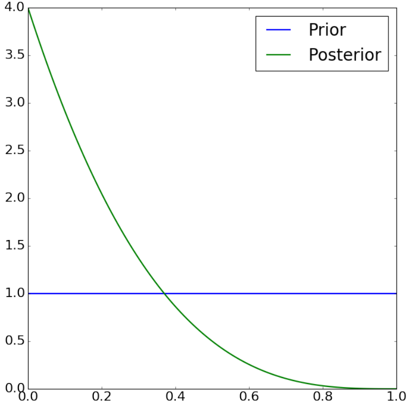
Coin Toss Example

Let us consider the Bernoulli model for a coin toss, for $\theta \in [0, 1]$

$$p(\mathsf{H} \mid \theta) = \theta$$

Suppose after three independent coin tosses, you get T, T, T. What is the maximum likelihood estimate for θ?

What is the posterior distribution over θ, assuming a uniform prior on θ?



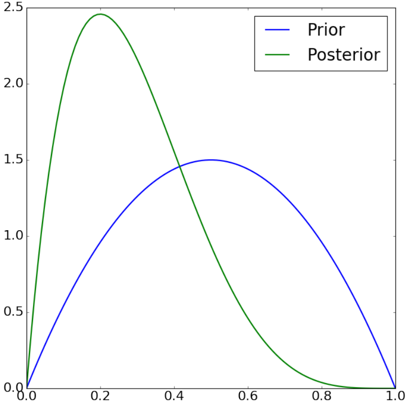
Coin Toss Example

Let us consider the Bernoulli model for a coin toss, for $\theta \in [0, 1]$

$$p(\mathsf{H} \mid \theta) = \theta$$

Suppose after three independent coin tosses, you get T, T, T. What is the maximum likelihood estimate for θ?

What is the posterior distribution over θ, assuming a $\mathrm{Beta}(2, 2)$ prior on θ?



Least Squares and MLE (Gaussian Noise)

Least Squares

Objective Function

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2$$

MLE (Gaussian Noise)

Likelihood

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left(-\frac{(y_i - \mathbf{w} \cdot \mathbf{x}_i)^2}{2\sigma^2}\right)$$

Least Squares and MLE (Gaussian Noise)

Least Squares

Objective Function

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2$$

MLE (Gaussian Noise)

Likelihood

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left(-\frac{(y_i - \mathbf{w} \cdot \mathbf{x}_i)^2}{2\sigma^2}\right)$$

For estimating $\mathbf{w}$, the negative log-likelihood under Gaussian noise has the same form as the least squares objective

Least Squares and MLE (Gaussian Noise)

Least Squares

Objective Function

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2$$

MLE (Gaussian Noise)

Likelihood

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left(-\frac{(y_i - \mathbf{w} \cdot \mathbf{x}_i)^2}{2\sigma^2}\right)$$

For estimating $\mathbf{w}$, the negative log-likelihood under Gaussian noise has the same form as the least squares objective

Alternatively, we can model the data (only y_i-s) as being generated from a distribution defined by exponentiating the negative of the objective function

What Data Model Produces the Ridge Objective?

We have the Ridge Regression Objective, let $\mathcal{D} = \langle(\mathbf{x}_i, y_i)\rangle_{i=1}^{N}$ denote the data

$$\mathcal{L}_{\text{ridge}}(\mathbf{w}; \mathcal{D}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^{\mathsf{T}}\mathbf{w}$$

What Data Model Produces the Ridge Objective?

We have the Ridge Regression Objective, let $\mathcal{D} = \langle(\mathbf{x}_i, y_i)\rangle_{i=1}^{N}$ denote the data

$$\mathcal{L}_{\text{ridge}}(\mathbf{w}; \mathcal{D}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^{\mathsf{T}}\mathbf{w}$$

Let's rewrite this objective slightly, scaling by $\frac{1}{2\sigma^2}$ and setting $\lambda = \frac{\sigma^2}{\tau^2}$. To avoid ambiguity, we'll denote this by $\widetilde{\mathcal{L}}$

$$\widetilde{\mathcal{L}}_{\text{ridge}}(\mathbf{w}; \mathcal{D}) = \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{1}{2\tau^2}\mathbf{w}^{\mathsf{T}}\mathbf{w}$$

What Data Model Produces the Ridge Objective?

We have the Ridge Regression Objective, let $\mathcal{D} = \langle(\mathbf{x}_i, y_i)\rangle_{i=1}^N$ denote the data

$$\mathcal{L}_{\text{ridge}}(\mathbf{w}; \mathcal{D}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^{\mathsf{T}}\mathbf{w}$$

Let's rewrite this objective slightly, scaling by $\frac{1}{2\sigma^2}$ and setting $\lambda = \frac{\sigma^2}{\tau^2}$. To avoid ambiguity, we'll denote this by $\widetilde{\mathcal{L}}$

$$\widetilde{\mathcal{L}}_{\text{ridge}}(\mathbf{w}; \mathcal{D}) = \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{1}{2\tau^2}\mathbf{w}^{\mathsf{T}}\mathbf{w}$$

Let $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_N$ and $\boldsymbol{\Lambda} = \tau^2 \mathbf{I}_D$, where $\mathbf{I}_m$ denotes the $m \times m$ identity matrix

$$\widetilde{\mathcal{L}}_{\text{ridge}}(\mathbf{w}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}\mathbf{w}$$

What Data Model Produces the Ridge Objective?

We have the Ridge Regression Objective, let $\mathcal{D} = \langle(\mathbf{x}_i, y_i)\rangle_{i=1}^{N}$ denote the data

$$\mathcal{L}_{\text{ridge}}(\mathbf{w}; \mathcal{D}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^{\mathsf{T}}\mathbf{w}$$

Let's rewrite this objective slightly, scaling by $\frac{1}{2\sigma^2}$ and setting $\lambda = \frac{\sigma^2}{\tau^2}$. To avoid ambiguity, we'll denote this by $\widetilde{\mathcal{L}}$

$$\widetilde{\mathcal{L}}_{\text{ridge}}(\mathbf{w}; \mathcal{D}) = \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{1}{2\tau^2}\mathbf{w}^{\mathsf{T}}\mathbf{w}$$

Let $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_N$ and $\boldsymbol{\Lambda} = \tau^2 \mathbf{I}_D$, where $\mathbf{I}_m$ denotes the $m \times m$ identity matrix

$$\widetilde{\mathcal{L}}_{\text{ridge}}(\mathbf{w}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}\mathbf{w}$$

Taking the negation of $\widetilde{\mathcal{L}}_{\text{ridge}}(\mathbf{w}; \mathcal{D})$ and exponentiating gives us a non-negative function of $\mathbf{w}$ and $\mathcal{D}$ which after normalisation gives a density function

$$f(\mathbf{w}; \mathcal{D}) = \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{w})\right) \cdot \exp\left(-\frac{1}{2}\mathbf{w}^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}\mathbf{w}\right)$$

Bayesian Linear Regression (and connections to Ridge)

Let's start with the form of the density function we had on the previous slide and factor it.

$$f(\mathbf{w}; \mathcal{D}) = \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{w})\right) \cdot \exp\left(-\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{\Lambda}^{-1}\mathbf{w}\right)$$

Bayesian Linear Regression (and connections to Ridge)

Let's start with the form of the density function we had on the previous slide and factor it.

$$f(\mathbf{w};\mathcal{D}) = \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{w})\right)\cdot\exp\left(-\frac{1}{2}\mathbf{w}^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}\mathbf{w}\right)$$

We'll treat σ as fixed and not as a parameter. Up to a constant factor (which does't matter when optimising w.r.t. $\mathbf{w}$), we can rewrite this as

$$\underbrace{p(\mathbf{w}\mid\mathbf{X},\mathbf{y})}_{\text{posterior}} \propto \underbrace{\mathcal{N}(\mathbf{y}\mid\mathbf{X}\mathbf{w},\Sigma)}_{\text{Likelihood}}\cdot\underbrace{\mathcal{N}(\mathbf{w}\mid\mathbf{0},\boldsymbol{\Lambda})}_{\text{prior}}$$

where $\mathcal{N}(\cdot\mid\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the density of the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

- What the ridge objective is actually finding is the maximum a posteriori or (MAP) estimate which is a mode of the posterior distribution
- The linear model is as described before with Gaussian noise
- The prior distribution on $\mathbf{w}$ is assumed to be a spherical Gaussian

Connections to Lasso

Similarly, the lasso objective finds MAP with Laplacian prior:

- Recall that $\text{Lap}(x; \mu, \gamma) = (1/2\gamma) \cdot \exp(-|x-\mu|/\gamma)$
- Lasso objective:

$$\mathcal{L}_{\text{lasso}}(\mathbf{w}; \mathcal{D}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \sum_{i=1}^{D} |w_i|$$

- Setting $\lambda = 4$, multiplying by $-1/2$, and exponentiating:

$$g(\mathbf{w}, \mathcal{D}) = \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{w})\right) \cdot \exp\left(-2 \cdot \sum_{i=1}^{D} |w_i|\right)$$

- Observe that

$$\exp\left(-2 \cdot \sum_{i=1}^{D} |w_i|\right) = \prod_{i=1}^{D} \exp(-2 \cdot |w_i|)$$

- That's a product of Laplacian distributions:
  $\text{Lap}(x; 0, 1/2) = \exp(-2 \cdot |x|)$

Full Bayesian Prediction

The posterior distribution over parameters $\mathbf{w}$ in the Bayesian approach is

$$\underbrace{p(\mathbf{w} \mid \mathbf{X}, \mathbf{y})}_{\text{posterior}} \propto \underbrace{p(\mathbf{y} \mid \mathbf{X}, \mathbf{w})}_{\text{likelihood}} \cdot \underbrace{p(\mathbf{w})}_{\text{prior}}$$

- If we use the MAP estimate, as we get more samples the posterior peaks at the MLE
- When, data is scarce rather than picking a single estimator (like MAP) we can sample from the full posterior

For $\mathbf{x}_{\text{new}}$, we can output the entire distribution over our prediction $\widehat{y}$ as

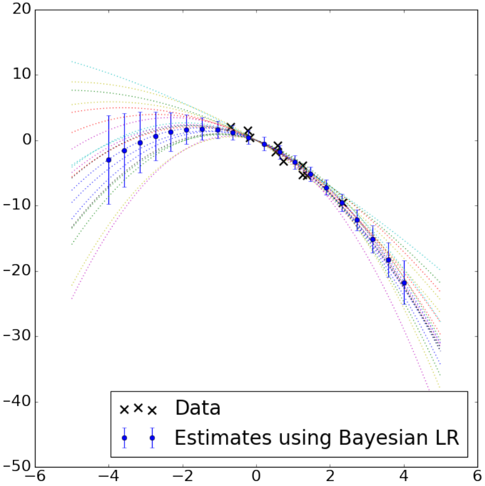
$$p(y \mid \mathcal{D}) = \int_{\mathbf{w}} \underbrace{p(y \mid \mathbf{w}, \mathbf{x}_{\text{new}})}_{\text{model}} \cdot \underbrace{p(\mathbf{w} \mid \mathcal{D})}_{\text{posterior}} \, \mathrm{d}\mathbf{w}$$

This integration is often computationally very hard!

Full Bayesian Approach for Linear Regression

For the linear model with Gaussian noise and a Gaussian prior on $\mathbf{w}$, the full Bayesian prediction distribution for a new point $\mathbf{x}_{\text{new}}$ can be expressed in closed form.

$$p(y \mid \mathcal{D}, \mathbf{x}_{\text{new}}, \sigma^2) = \mathcal{N}(\mathbf{w}_{\text{map}}^{\mathsf{T}} \mathbf{x}_{\text{new}}, (\sigma(\mathbf{x}_{\text{new}}))^2)$$



See Murphy Sec 7.6 for calculations

Remarks on Prior Distribution

- Presence of prior point of criticism in Bayesian approach
- Prior should incorporate all reasonable background information (e.g. domain-specific information, previous knowledge)
- If no background information available choose non-informative prior (uniform over expected range of possible values)
- Conjugate priors allow for analytical solutions
- Bernstein-von Mises Theorem: For sufficiently large sample size, posterior distribution becomes independent of prior distribution

Remarks on Prior Distribution

- Presence of prior point of criticism in Bayesian approach
- Prior should incorporate all reasonable background information (e.g. domain-specific information, previous knowledge)
- If no background information available choose non-informative prior (uniform over expected range of possible values)
- Conjugate priors allow for analytical solutions
- Bernstein-von Mises Theorem: For sufficiently large sample size, posterior distribution becomes independent of prior distribution (terms and conditions apply)

Summary : Bayesian Machine Learning

In the Bayesian view, in addition to modelling the output y as a random variable given the parameters $\mathbf{w}$ and input $\mathbf{x}$, we also encode prior belief about the parameters $\mathbf{w}$ as a probability distribution $p(\mathbf{w})$.

- If the prior has a parametric form, they are called hyperparameters
- The posterior over the parameters $\mathbf{w}$ is updated given data
- Either pick point (plugin) estimates, *e.g.*, maximum a posteriori
- Or as in the full Bayesian approach use the entire posterior to make prediction (this is often computationally intractable)
- Choice of prior can be difficult?