

Machine Learning Basics: Bayesian Statistics

Sargur N. Srihari
srihari@cedar.buffalo.edu

This is part of lecture slides on [Deep Learning](http://www.cedar.buffalo.edu/~srihari/CSE676):
<http://www.cedar.buffalo.edu/~srihari/CSE676>

Topics in Machine Learning Basics

1. Learning Algorithms
2. Capacity, Overfitting and Underfitting
3. Hyperparameters and Validation Sets
4. Estimators, Bias and Variance
5. Maximum Likelihood Estimation
6. Bayesian Statistics
7. Supervised Learning Algorithms
8. Unsupervised Learning Algorithms
9. Stochastic Gradient Descent
10. Building a Machine Learning Algorithm
11. Challenges Motivating Deep Learning

Topics in Bayesian Statistics

- Frequentist versus Bayesian Statistics
- Prior Probability Distribution
- Bayesian Estimation
- Ex: Bayesian Linear Regression
- Maximum A Posteriori (MAP) Estimation

Frequentist Statistics

- So far we have discussed *frequentist statistics*
- The approaches are based on estimating a single value of θ , then making all predictions thereafter based on that one estimate
- Summary of the frequentist perspective:
 - True parameter value θ is fixed but unknown
 - Point estimate $\hat{\theta}$ is a random variable
 - On account of it being a function of the dataset
- The Bayesian perspective is quite different

Bayesian Perspective

- The Bayesian approach is to consider all possible values of θ before making predictions
- The Bayesian uses probability to reflect degrees of uncertainty in states of knowledge
- The dataset is directly observed and so is not random
- On the other hand, the true parameter θ is unknown or uncertain and is thus represented as a random variable

Prior Distribution

- Before observing the data, we represent our knowledge of θ is using the prior probability distribution $p(\theta)$
 - Sometimes simply referred to as the prior
 - ML practitioner selects prior distribution to be broad
 - i.e., with high entropy to reflect high uncertainty
- Examples
 - θ is in a finite range/volume with uniform distribution
 - Many priors reflect preferences for “simpler” solutions
 - Such as smaller magnitude coefficients
 - Or a function closer to being constant

Bayes rule in Bayesian approach

- Consider we have a set of samples $\{x^{(1)}, \dots, x^{(m)}\}$
- We can recover the effect of data on our belief about θ by combining the data likelihood $p(x^{(1)}, \dots, x^{(m)} | \theta)$ with the prior $p(\theta)$ via Bayes rule:

$$p(\theta | x^{(1)}, \dots, x^{(m)}) = \frac{p(x^{(1)}, \dots, x^{(m)} | \theta) p(\theta)}{p(x^{(1)}, \dots, x^{(m)})}$$

From prior to posterior

- In Bayesian estimation the prior begins as a relatively uniform or Gaussian with high entropy
- Data causes the posterior to lose entropy
 - And concentrate around a few highly likely values of parameters

Two Differences between maximum likelihood and Bayesian estimation

1. Making predictions
2. Contribution of the prior distribution

1. Making Predictions in Bayesian approach

- MLE predictions use point estimate θ
- Bayesian predicts using full distribution over θ
 - Ex: after m samples, prediction over sample x^{m+1} is

$$p(x^{(m+1)} | x^{(1)}, \dots, x^{(m)}) = \int p(x^{(m+1)} | \theta) p(\theta | x^{(1)}, \dots, x^{(m)}) d\theta$$

- Where

$$p(\theta | x^{(1)}, \dots, x^{(m)}) = \frac{p(x^{(1)}, \dots, x^{(m)} | \theta) p(\theta)}{p(x^{(1)}, \dots, x^{(m)})}$$

- Here each value of θ with positive probability density contributes to the prediction of the next example
 - With the contribution weighted by the posterior density itself
- After having observed $\{x^{(1)}, \dots, x^{(m)}\}$, if we are still quite uncertain about the value of θ , then the uncertainty is incorporated directly into any predictions we might make

Handling uncertainty in θ

- In the frequentist approach, uncertainty in a given point estimate of θ is handled by evaluating its variance
 - Variance of the estimator is an assessment of how the estimate might change with alternative samplings of the observed data
- Bayesian answer to the question of how to handle uncertainty in the estimator is to simply integrate over it

Bayesian approach avoids overfitting

- Bayesian approach tends to protect well against overfitting
 - In the Bayesian approach there is no fitting, just computing the posterior from the prior
- Integral is just an application of the laws of probability making the Bayesian approach simple to justify
- While the frequentist machinery for constructing an estimator is based on an ad-hoc decision to summarize all knowledge contained in the dataset with a single point estimate

Second difference: Bayesian vs MLE

2. Contribution of the prior distribution

- Prior has influence of shifting probability mass function towards regions of the parameter space that are preferred a priori
- In practice prior expresses a preference over models that are simpler or more smooth
- Critics of Bayesian approach identify the prior as a source of subjective human judgment affecting the predictions

When is Bayesian approach better?

- Bayesian methods typically generalize much better when limited training data is available
- But suffer from high computational cost when the number of training examples is large

Ex: Bayesian Linear Regression

- Consider the Bayesian estimation approach to learning linear regression parameters
- Here we learn a linear mapping from an input vector $\mathbf{x} \in \mathbb{R}^n$ to predict value of a scalar $y \in \mathbb{R}$
- The prediction is parameterized by the vector $\mathbf{w} \in \mathbb{R}^n$:
$$\hat{y} = \mathbf{w}^T \mathbf{x}$$
- Given a set of m training samples $(\mathbf{X}^{(\text{train})}, \mathbf{y}^{(\text{train})})$,
 - we can express the prediction of y over the entire training set as
$$\hat{\mathbf{y}}^{(\text{train})} = \mathbf{X}^{(\text{train})} \mathbf{w}$$

Prediction as a Gaussian conditional

- Prediction $\mathbf{y}^{(\text{train})}$ can be expressed as a Gaussian conditional distribution:

$$p(\mathbf{y}^{(\text{train})} \mid \mathbf{X}^{(\text{train})}, \mathbf{w}) = \mathcal{N}(\mathbf{y}^{(\text{train})}; \mathbf{X}^{(\text{train})}\mathbf{w}, \mathbf{I}) \\ \propto \exp\left(-\frac{1}{2}(\mathbf{y}^{(\text{train})} - \mathbf{X}^{(\text{train})}\mathbf{w})^\top (\mathbf{y}^{(\text{train})} - \mathbf{X}^{(\text{train})}\mathbf{w})\right)$$

- Where we follow the standard MSE formulation in assuming that the Gaussian variance on y is one
- In what follows we refer to $(\mathbf{X}^{(\text{train})}, \mathbf{y}^{(\text{train})})$ as simply (\mathbf{X}, \mathbf{y})

Gaussian prior distribution

- To specify a posterior distribution over the parameters \mathbf{w} , we first need to specify a prior
- The prior should reflect our naive belief about the value of these parameters
 - Assume a fairly broad distribution to express high degree of uncertainty about θ
- For real-valued parameters it is common to use a Gaussian as a prior distribution

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0) \propto \exp \left(-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0^{-1}(\mathbf{w} - \boldsymbol{\mu}_0) \right)$$

- where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Lambda}_0$ are prior distribution mean and covariance matrix; typically assume diagonal $\boldsymbol{\Lambda}_0 = \text{diag}(\lambda_0)$

Posterior on the weights

- With prior specified, we can determine the posterior distribution over parameters

$$\begin{aligned}
 p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}) &\propto p(\mathbf{y} \mid \mathbf{X}, \mathbf{w})p(\mathbf{w}) \\
 &\propto \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w})\right) \exp\left(-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0^{-1}(\mathbf{w} - \boldsymbol{\mu}_0)\right) \\
 &\propto \exp\left(-\frac{1}{2}\left(-2\mathbf{y}^\top \mathbf{X}\mathbf{w} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} + \mathbf{w}^\top \boldsymbol{\Lambda}_0^{-1}\mathbf{w} - 2\boldsymbol{\mu}_0^\top \boldsymbol{\Lambda}_0^{-1}\mathbf{w}\right)\right)
 \end{aligned}$$

Now define $\boxed{\boldsymbol{\Lambda}_m = (\mathbf{X}^\top \mathbf{X} + \boldsymbol{\Lambda}_0^{-1})^{-1}}$ and $\boxed{\boldsymbol{\mu}_m = \boldsymbol{\Lambda}_m (\mathbf{X}^\top \mathbf{y} + \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0)}$

- Using these new variables, we find that the posterior can be written as a Gaussian distribution:

$$\begin{aligned}
 p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}) &\propto \exp\left(-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_m)^\top \boldsymbol{\Lambda}_m^{-1}(\mathbf{w} - \boldsymbol{\mu}_m) + \frac{1}{2}\boldsymbol{\mu}_m^\top \boldsymbol{\Lambda}_m^{-1}\boldsymbol{\mu}_m\right) \\
 &\propto \exp\left(-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_m)^\top \boldsymbol{\Lambda}_m^{-1}(\mathbf{w} - \boldsymbol{\mu}_m)\right).
 \end{aligned}$$

All terms that do not include the parameter vector \mathbf{w} have been omitted. They are implied by normalization.

Maximum *A Posteriori* (MAP) Estimation

- Most principled approach is to use full posterior distribution over parameters θ , it is still often desirable to have a single point estimate
- Most often a Bayesian posterior is intractable
- A point estimate offers a tractable solution
- Rather than simply returning the maximum likelihood estimate, we can still gain some benefit of the Bayesian approach by allowing the prior to influence the choice of the point estimate

Intuition on Bayesian inference

- In most situations we set $\mu_0=0$
- If we set $\Lambda_0=(1/\alpha)\mathbf{I}$ then μ_m gives the same estimate of \mathbf{w} as does frequentist linear regression with a weight decay penalty of $\alpha \mathbf{w}^T \mathbf{w}$
- The most important difference is that the Bayesian estimate provides a covariance matrix, showing how likely all the different values of \mathbf{w} are, rather than providing only the estimate μ_m

The MAP point estimate

- The MAP estimate chooses the point of maximum a posteriori probability (or maximum probability density in the case of a continuous distribution)

$$\theta_{\text{MAP}} = \arg \max_{\theta} p(\theta | x) = \arg \max_{\theta} \log p(x | \theta) + \log p(\theta)$$

- In the rhs $\log p(x|\theta)$ is the standard log-likelihood term
- $\log p(\theta)$ corresponds to the prior distribution

Example of MAP estimate

- Linear regression with Gaussian prior on w
- If this prior is given by $\mathcal{N}(w; \mathbf{0}, (1/\lambda)\mathbf{I}^2)$ then the log prior term is proportional to the familiar $\lambda w^T w$ weight decay penalty plus a term that does not depend on w and does not affect the learning process
- MAP Bayesian inference with a Gaussian prior on the weights thus corresponds to weight decay

Advantage of MAP point estimate

- Leverages information brought by the prior and cannot be found in the training data
- It helps reduce the variance in the MAP point estimate (in comparison to the ML estimate)
- However it does so at the risk of increased bias
- A more complicated penalty term can be derived by using a mixture of Gaussians rather than a single Gaussian distribution as prior