TDA231
Going Bayesian

Devdatt Dubhashi
dubhashi@chalmers.se

Dept. of Computer Science and Engg.
Chalmers University

March 26, 2018

Introduction

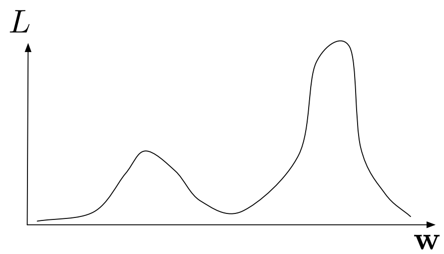
- We have seen two ways of finding the 'best' parameter values:
  - Those that minimise the *loss*.
  - Those that maximise the *likelihood*.
  - If noise is Gaussian, both are the same:

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

- Is this the 'right' set of parameters?
- Is there a 'right' set of parameters?

Introduction
D. Dubhashi
Introduction
Bayesian machine learning
Example
Marginal likelihood
Choosing a prior
Summary

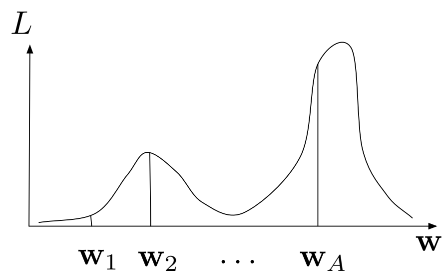
Problems with a point estimate



- Might be more than one 'best' value.
- Might not be a single representative value.
- Different values might give very different predictions.
- Is there an alternative?

Introduction
D. Dubhashi
Introduction
Bayesian machine learning
Example
Marginal likelihood
Choosing a prior
Summary

Averaging



- Prediction is some function of $\mathbf{w}$. Say $f(\mathbf{w})$.
- Choose A different values – $\mathbf{w}_1, \ldots, \mathbf{w}_A$.
- Compute $\sum_{a=1}^{A} q_a f(\mathbf{w}_a)$
- q_a is proportional to L (subject to $\sum_a q_a = 1$)
- Increasing A seems like a good idea....

Introduction
D. Dubhashi
Introduction
Bayesian machine learning
Example
Marginal likelihood
Choosing a prior
Summary

Example

- Olympic 100 m data.
- Want to predict winning time at London 2012 – t_{new}.
- Choose 2 'good' values of $\mathbf{w}$
  - $\mathbf{w}_1$ predicts $t_{\text{new}} = 9.5\ s$
  - $\mathbf{w}_2$ predicts $t_{\text{new}} = 9.2\ s$
- According to likelihood, $\mathbf{w}_2$ is twice as likely as $\mathbf{w}_1$.
  - $q_1 + q_2 = 1,\ q_2 = 2q_1$.
  - Therefore: $q_1 = 1/3,\ q_2 = 2/3$
- Average prediction is $(1/3) \times 9.5 + (2/3) \times 9.2 = 9.3$

Bayes rule

Introduction
D. Dubhashi

- Bayes rule:

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

Bayes rule

Introduction
D. Dubhashi

- Bayes rule:

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

- **Posterior density**: $p(\mathbf{w}|\mathbf{X},\mathbf{t})$
  - This is what we're after.

Bayes rule

Introduction
D. Dubhashi

- Bayes rule:

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

- **Posterior density**: $p(\mathbf{w}|\mathbf{X},\mathbf{t})$
  - This is what we're after.
- Likelihood : $p(\mathbf{t}|\mathbf{X},\mathbf{w})$
  - We've used this before.

Bayes rule

- Bayes rule:

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

- **Posterior density**: $p(\mathbf{w}|\mathbf{X},\mathbf{t})$
  - This is what we're after.
- Likelihood : $p(\mathbf{t}|\mathbf{X},\mathbf{w})$
  - We've used this before.
- Prior density: $p(\mathbf{w})$
  - This is new: do we know anything about the parameters before we see any data?

Bayes rule

- Bayes rule:

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

- **Posterior density**: $p(\mathbf{w}|\mathbf{X},\mathbf{t})$
  - This is what we're after.
- Likelihood : $p(\mathbf{t}|\mathbf{X},\mathbf{w})$
  - We've used this before.
- Prior density: $p(\mathbf{w})$
  - This is new: do we know anything about the parameters before we see any data?
- Marginal likelihood: $p(\mathbf{t}|\mathbf{X})$
  - This is new: $\mathbf{w}$ isn't in here. It is a normalisation constant. Ensures $\int p(\mathbf{w}|\mathbf{X},\mathbf{t})\ d\mathbf{w} = 1$.

Computing the posterior

- Unfortunately, computing the posterior is hard...
- ...because marginal likelihood $p(\mathbf{t}|\mathbf{X})$ is hard to compute:

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

Computing the posterior

- Unfortunately, computing the posterior is hard...
- ...because marginal likelihood $p(\mathbf{t}|\mathbf{X})$ is hard to compute:

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

- In some cases we can do it (this lecture).
- In most we can't and are forced to (later in course):
  - Approximate $p(\mathbf{w}|\mathbf{X},\mathbf{t})$ with something else.
  - Sample from $p(\mathbf{w}|\mathbf{X},\mathbf{t})$ (incredibly, we can sample from it even if we can't compute it!)

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, e.g. http://en.wikipedia.org/wiki/Conjugate_prior

Why is this important?

- Bayes rule:

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

- If prior and likelihood are conjugate, we **know** the form of $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$
- Therefore, we **know** the form of the normalising constant.
- Therefore, we **don't need** to compute $p(\mathbf{t}|\mathbf{X})$

Why is this important?

- Bayes rule:

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

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

Example - Olympic data

- We'll use the (Gaussian) likelihood we used for maximum likelihood:

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

Example - Olympic data

- We'll use the (Gaussian) likelihood we used for maximum likelihood:

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

- The prior conjugate to the Gaussian is Gaussian. So:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S}), \ \mathbf{S} = \begin{bmatrix} 100 & 0 \\ 0 & 5 \end{bmatrix}$$

  - Mean ($\mathbf{0}$) and covariance ($\mathbf{S}$) are design choices.

Example - Olympic data

- We'll use the (Gaussian) likelihood we used for maximum likelihood:

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

- The prior conjugate to the Gaussian is Gaussian. So:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S}), \ \mathbf{S} = \begin{bmatrix} 100 & 0 \\ 0 & 5 \end{bmatrix}$$

  - Mean ($\mathbf{0}$) and covariance ($\mathbf{S}$) are design choices.

- Posterior **must be** gaussian with unknown parameters:

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

Finding posterior parameters

- Ignoring normalising constant, the posterior is:

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

Finding posterior parameters

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

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

- Posterior (from previous slide):

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

Finding posterior parameters

- Equate individual terms on each side.
- Covariance:

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

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

- Mean:

$$2\mathbf{w}^\mathsf{T}\mathbf{\Sigma}^{-1}\boldsymbol{\mu} = \frac{2}{\sigma^2}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{t}$$

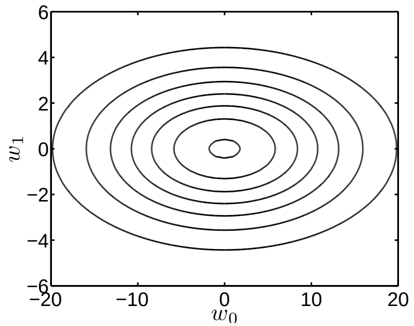
$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\mathbf{\Sigma}\mathbf{X}^\mathsf{T}\mathbf{t}$$

Olympic example

- To make numbers better, rescape olympic year:
  - $1896 = 1, 1900 = 2, \ldots, 2008 = 27, 2012 = 28$

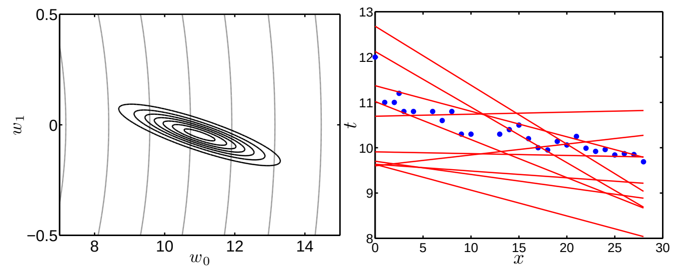
Olympic example

- To make numbers better, rescape olympic year:
  - $1896 = 1, 1900 = 2, \ldots, 2008 = 27, 2012 = 28$
- Prior density:



- Mean ($\mathbf{0}$) and covariance ($\mathbf{S}$).
- Quite a *vague* prior.

Olympic example



Posterior (left) (prior shown in grey, zoomed in) and functions corresponding to some $\mathbf{w}$ sampled from posterior (right).

Olympic example – predictions

- Our motivation for being Bayesian was to be able to average predictions (at $\mathbf{w}_{\text{new}}$) over all $\mathbf{w}$
- We have the full posterior distribution over all possible values of $\mathbf{w}$, it is also Gaussian and we computed the parameters.

Olympic example – predictions

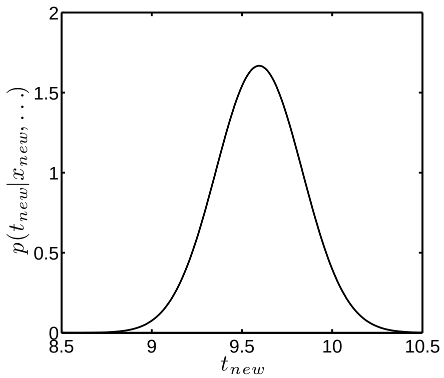
Introduction
D. Dubhashi
Introduction
Bayesian machine learning
Example
Marginal likelihood
Choosing a prior
Summary

- Our motivation for being Bayesian was to be able to average predictions (at $\mathbf{w}_{\text{new}}$) over all $\mathbf{w}$
- We have the full posterior distribution over all possible values of $\mathbf{w}$, it is also Gaussian and we computed the parameters.
- We can compute exactly, the predictive density to make probabilistic predictions:

$$p(t_{\text{new}}|\mathbf{X},\mathbf{t},\mathbf{x}_{\text{new}},\sigma^2) = \mathcal{N}(\mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\mu}, \sigma^2 + \mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{x}_{\text{new}})$$

Olympic example – predictions

Introduction
D. Dubhashi
Introduction
Bayesian machine learning
Example
Marginal likelihood
Choosing a prior
Summary



Predictive density at 2012 Olympics. Note that σ^2 was fixed at 0.05.

Computing posterior: recipe

Introduction
D. Dubhashi
Introduction
Bayesian machine learning
Example
Marginal likelihood
Choosing a prior
Summary

- (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 $p(\mathbf{t}|\mathbf{X},\sigma^2)$, the normalising thing in Bayes rule.
- We stated that it was equal to (because it's a normalising thing):

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

- We're averaging over all values of $\mathbf{w}$ to get a value for **how good the model is**.
  - How likely is $\mathbf{t}$ given $\mathbf{X}$ and the model. e.g. 'first order polynomial'.
- Can use this to compare models.

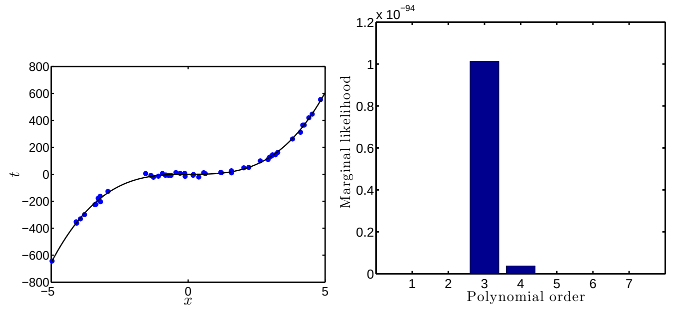
Marginal likelihood

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

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

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

Marginal likelihood – example



Some data generated from a 3rd order polynomial (left) and the marginal likelihood for polynomials of varying order.

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.

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.

Choosing a prior

Introduction
D. Dubhashi
Introduction
Bayesian machine learning
Example
Marginal likelihood
Choosing a prior
Summary

- 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'

Choosing a prior

Introduction
D. Dubhashi
Introduction
Bayesian machine learning
Example
Marginal likelihood
Choosing a prior
Summary

- 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!)

Choosing a prior

Introduction
D. Dubhashi
Introduction
Bayesian machine learning
Example
Marginal likelihood
Choosing a prior
Summary

- 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).

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*.

Introduction
D. Dubhashi
Introduction
Bayesian machine learning
Example
Marginal likelihood
Choosing a prior
Summary

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.

Introduction
D. Dubhashi
Introduction
Bayesian machine learning
Example
Marginal likelihood
Choosing a prior
Summary