MAP 534
Introduction to machine learning

Bayesian machine learning

Alain Durmus

Outline

Motivations

Bayesian statistics

- Decide what the input-output pairs are.
- Decide how to encode inputs and outputs. This defines the input space X, and the output space Y and the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$.
- Choose a class of hypotheses/representations $\mathcal{F} = \{f_w : \mathrm{X} \to \mathrm{Y} \ : \ w \in \mathbb{R}^d\}$.
- Choose a loss function ℓ.
- Define the error function

$$E_\ell(w) = N^{-1} \sum_{i=1}^{N} \ell(y_i, f_w(x_i)) \ . \tag{1}$$

- Choose an algorithm to solve

$$\text{minimize } E_\ell \ . \tag{2}$$

- How to do so: vanish the gradient or gradient descent (see later)...

Modeling view of machine learning

- Decide what the input-output pairs are.
- Decide how to encode inputs and outputs. This defines the input space X, and the output space Y and $\mathcal{D}$.
- Choose a class of hypotheses/representations $\mathcal{F} = \{f_w : \mathrm{X} \to \mathrm{Y} \ : \ w \in \mathbb{R}^d\}$.
- Choose a loss function ℓ.
- Define the error function

$$E_\ell(w) = N^{-1} \sum_{i=1}^{n} \ell(y_i, f_w(x_i)) \ . \tag{3}$$

- Choose an algorithm to solve

$$\hat{w} \in \operatorname{argmin} E_\ell \ . \tag{4}$$

- Prediction:

$$\hat{y}_{\mathbf{pred}} = f_{\hat{w}}(x_{\mathbf{new}}) \ . \tag{5}$$

- This framework is called the deterministic discriminative setting.

Going back to the example of polynomial curve fitting

- Consider N observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ such that

$$x_i \in \mathsf{X} = [0,1] \text{ and } y_i \in \mathsf{Y} = \mathbb{R} \,. \tag{6}$$

- We consider here that

$$\mathcal{P}_d = \left\{ f_w(x) = \sum_{i=1}^{d} w_i x^{(i)} \right\} \,. \tag{7}$$

- This corresponds to the choice of basis function $\phi_j(x) = x^j$.
- Justification: polynomials can approximate continuous any function on $[0,1]$.
- LSE:

$$\hat{w} \in \operatorname*{argmin}_{w} E(w) \,, \quad E(w) = \frac{1}{2N} \sum_{j=1}^{N} \left\{ y_j - \sum_{i=1}^{d} w_i x_j^i \right\}^2 \,. \tag{8}$$

- Two main questions still remain:
  - Can we weight the possible choices for $\mathcal{F}/d$?
  - Can we quantify the uncertainty of the prediction?
- The two questions are related and addressed with the use of Bayesian statistics:
  - For these two problems, we give some weights/probabilities on models/coefficients based on a priori knowledge.
  - Regarding the second point, Bayesian inference “sees” the parameter w as random!

Probabilistic modeling in machine learning

- No modeling view discussed previously: no probability!
- Here we consider a statistical model on the observations $\{(x_i, y_i\}_{i=1}^N$.
- This model as in your statistics course is specified by a likelihood, i.e., a family of parametrized probability density functions (p.d.f.)

$$\{(x, y) \mapsto \mathrm{L}_w(x, y) \,:\, w \in \Theta \subset \mathbb{R}^d\} \,.$$

- Examples:
  - Regression:

$$Y_i = f_w(X_i) + \epsilon_i \,, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1) \,. \tag{9}$$

Likelihood:

$$\mathrm{L}_w(x, y) = ? \tag{10}$$

- No modeling view discussed previously: no probability!
- Here we consider a statistical model on the observations $\{(x_i, y_i\}_{i=1}^N$.
- This model as in your statistics course is specified by a likelihood, i.e., a family of parametrized probability density functions (p.d.f.)

$$\{(x, y) \mapsto \mathrm{L}_w(x, y) \ : \ w \in \Theta \subset \mathbb{R}^d\} \ .$$

- Examples:
  - Classification ($\mathsf{Y} = \{0, 1\}$):

$$Y_i = \mathbb{1}\{f_w(X_i) + \epsilon_i \geq 0\} \ , \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1) \ . \tag{11}$$

    Likelihood:

$$\mathrm{L}_w(x, y) = ? \tag{12}$$

Probabilistic modeling in machine learning: discriminative setting

- No modeling view discussed previously: no probability!
- Here we consider a statistical model on the observations $\{(x_i, y_i)\}_{i=1}^N$.
- Here, we fix a distribution p for x but does not matter, so the likelihood has the form:

$$\left\{(x, y) \mapsto \mathrm{L}_w(x, y) = p_w(y|x)p(x) \,:\, w \in \Theta \subset \mathbb{R}^d\right\}.$$

- MLE:

$$\hat{w} \in \operatorname*{argmax}_w \{\log p_w(y|x)\} \tag{13}$$

- Prediction:

$$y_{\textbf{pred}} = \operatorname*{argmax}_y p_{\hat{w}}(y|x_{\textbf{new}}). \tag{14}$$

- In our examples, what would $y_{\textbf{pred}}$ be?
- We only care about the conditional $y|x$!
- This framework is referred to as the probabilistic discriminative framework.

Probabilistic modeling in machine learning: generative setting

- Here we consider a statistical model on the observations $\{(x_i, y_i)\}_{i=1}^{N}$.
- We can also take a generative model approach:

$$\{(x, y) \mapsto \mathrm{L}_w(x, y) = p_w(x|y)p(y) \,:\, w \in \Theta \subset \mathbb{R}^d\} \,.$$

- Still the MLE:

$$\hat{w} \in \underset{w}{\operatorname{argmax}}\{\log p_w(y|x)\} \tag{15}$$

- Prediction:

$$y_{\mathbf{pred}} = \underset{y}{\operatorname{argmax}}\, p_{\hat{w}}(y|x_{\mathbf{new}}) \,. \tag{16}$$

- What is $p_w(y|x)$? $p(y)$?
- Answer: Bayes theorem/formula and $p(y)$ is a prior to choose (details further...)
- This framework is referred to as the probabilistic generative framework.
- Pros: access to $p_{\hat{w}}$ which allows detection of outliers.
- Cons: computational demanding...
- Example in the next course!

Link with statistics: Gaussian regression and maximum likelihood estimation

- The statistical model associated with the LSE is:

$$Y_i \overset{\text{iid}}{\sim} \sum_{i=1}^{d} w_j \phi_j(X_i) + \sigma^2 Z_i \ , \quad i \in \{1, \ldots, N\} \ , \tag{17}$$

  where
  - $Z_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$;
  - X_i are i.i.d random variables with unknown distribution;
  - w is the parameter to infer.

- Then, if we are just interested in inferring w, the log-likelihood is

$$\ell(w) = \frac{1}{2\sigma} \sum_{i=1}^{N} \left\{ y_i - \sum_{i=1}^{d} w_j \phi_j(x_i) \right\}^2 \ , \tag{18}$$

  where the observations are $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$.

- Therefore, maximizing the log-likelihood leads to the same solution as minimizing the error function.

Uncertainty and model selection in frequentist statistics

- MLE: Only point estimate! No uncertainty quantification!
- Confident intervals on the coefficients of w defined by bootstrap:

$$\mathcal{D}_{\mathbf{rand},i} \subset \mathcal{D}_{\mathbf{train}} \text{ uniformly random }, \quad \text{ and consider } \hat{w}(d, \mathcal{D}_{\mathbf{rand},i}) \, . \quad (19)$$

- Consider the intervals which contains ..% of the solutions.
- Analysis of variance (ANOVA, using an F-test): test the null hypothesis that a model $\mathcal{M}_1$ is sufficient to explain the data against the alternative hypothesis that a more complex model $\mathcal{M}_2$.
- Do not generalize well and have pathologies[1]:
  - counter-intuitive behavior on confident intervals for some simple models;
  - p-values tend to overstate the evidence against the null no matter how large the sample size;
  - p-values is sensible to slight changes in the statistical models;
  - many frequentist methods regarding uncertainties/model selection does not follow likelihood principle: are based on hypothetical future observations.
- All these problems can be addressed by Bayesian inference!

[1] [Mur13, Section 6.6]

Going back to our questions

- Consider N observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ such that

$$x_i \in \mathsf{X} = [0, 1] \text{ and } y_i \in \mathsf{Y} = \mathbb{R} \; . \tag{20}$$

- We consider here that

$$\mathcal{P}_d = \left\{ f_w(x) = \sum_{i=1}^{d} w_i x^i \right\} \; . \tag{21}$$

- We would like to quantify uncertainties with respect to
    - the parameters w;
    - our prediction;
    - the hypothesis class complexity d.
- This generalizes to any parametrized hypothesis class

$$\mathcal{F} = \{ f_w : \mathsf{X} \to \mathsf{Y} \; : \; w \in \Theta \} \; . \tag{22}$$

Open questions

- Do we think that all hypothesis/models are equally probable... before we see any data?
- Here an hypothesis is a fixed function f_w for some fixed parameter w.
- What does the probability of a model/hypothesis even mean?
- Do we need to choose a single "best" model $\mathcal{F}$ or can we consider several $\mathcal{F}_1, \mathcal{F}_2$ for our predictions?
- We need a framework to answer such questions.

Bayes paradigm

- Bayes rule tells us how to do inference about hypothesis (the uncertain quantities) from data (measured quantities).
- Learning and prediction can be seen as a form of inference

$$p(\text{hypothesis}|\text{data}) = \frac{p(\text{data}|\text{hypothesis})p(\text{hypothesis})}{p(\text{data})} . \qquad (23)$$

- $p(\text{data}|\text{hypothesis})$ is the likelihood associated with the family of hypothesis we first consider.
- $p(\text{hypothesis}|\text{data})$ is called the posterior distribution of the hypothesis.
- However, in contrast to frequentist statistics, we choose a prior on our hypothesis!

- Bayes inference recipe:
    - Consider a staistical model for $\mathcal{D}$ parametrized by $w \in \Theta$:
    $$p(\mathcal{D}|w) = \mathrm{L}_w(x, y) \ . \tag{24}$$
    - We treat the likelihood as the conditional distribution of the data given the parameter!
    - Choose a prior for w, $p(w)$.
    - Consider the posterior:
    $$p(w|\mathcal{D}) \propto p(\mathcal{D}|w)p(w) = \mathrm{L}_w(x, y)p(w) \ . \tag{25}$$
    - All the conclusions are then drawn from the posterior.

- Consider the observations $\{y_i\}_{i=1}^{N}$ be i.i.d from

$$Y_i = \mathrm{Ber}(q)\ ,\ \ q \in [0,1] \text{ is the parameter to infer .} \tag{26}$$

- The likelihood is

$$p_q(y) = ? \tag{27}$$

- We chose as a prior Beta(α, β), for $\alpha, \beta > 0$:

$$p(q|y) \propto ? \tag{28}$$

- The posterior distribution for q is ?...
- Notation here if $p(q|y)$ is a conditional density:

$$p(q|y) \propto h(q,y)\ ,\ \text{ if } p(q|y) = h(q,y) \Big/ \int h(q,y)\mathrm{d}q\ \ . \tag{29}$$

- Bayes inference recipe:
  - Consider a staistical model for $\mathcal{D}$ parametrized by $w \in \Theta$:
$$p(\mathcal{D}|w) = \mathrm{L}_w(x, y) \ . \tag{30}$$
  - We treat the likelihood as the conditional distribution of the data given the parameter!
  - Choose a prior for w, $p(w)$.
  - Consider the posterior:
$$p(w|\mathcal{D}) \propto p(\mathcal{D}|w)p(w) = \mathrm{L}_w(x, y)p(w) \ . \tag{31}$$
  - All the conclusions are then drawn from the posterior.
  - The posterior is known up to a multiplicative constant:
$$\mathrm{Z}(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)\mathrm{d}w = \int \mathrm{L}_w(x, y)p(w)\mathrm{d}w \ . \tag{32}$$
    This constant is also known as the marginal likelihood.
  - In many models, this constant can not be computed and the posterior does not belong to "common" distribution.

Why Bayesian inference seems to be a sensible option?

- A robot, in order to behave intelligently, should be able to represent beliefs about propositions in the world:
  - charging station is at location (x,y,z)
  - that cat is hostile...
- Using probabilistic models, we want to represent the strengths of these beliefs, and be able to manipulate these beliefs based on a priori.
- The prior distribution models this prior knowledge.
- Data are then used to update our knowledge and give the posterior.
- Probabilistic learning can also be used for calibrated models and prediction uncertainty - getting systems that know what they do not know.

Prior and multiple explanations of the data

- Choosing a prior and following the Bayesian paradigm, we do not believe all models are equally probable to explain the data.
- We may believe that a simpler model is more probable than a complex one based on Occam's razzor (Aristotle, Ockham, Newton, Russel...)

  We consider it a good principle to explain the phenomena by the simplest hypothesis possible.

  - Ptolemy (c. AD 90 - c. 168) -
- Bayesian allows us to consider/combine a collection of hypothesis/models:
  - We do not know what particular function generated the data.
  - More than one of our models can perfectly fit the data.
  - We believe more than one of our models could have generated the data.
  - We want to reason in terms of a set of possible explanations, not just one.

Point estimate and uncertainty quantification on the parameter

- The first Bayesian estimator, the maximum a posterior estimator (MAP):

$$\hat{w}_{\mathsf{MAP}} \in \operatorname{argmax} p(w|\mathcal{D}) \; . \tag{33}$$

- The MAP is not fully Bayesian (not an admissible estimator)...
- The usual Bayesian estimator is the posterior mean:

$$\hat{w}_{\mathbf{post}} = \int w p(w|\mathcal{D}) \mathrm{d}w \; . \tag{34}$$

- To quantify the uncertainties over w we consider $1-\alpha$-credible region for $\alpha \in (0,1)$.
- C_α is set to be a $1-\alpha$-credible region if

$$\int \mathbb{1}\left\{w \in \mathsf{C}_\alpha\right\} p(w|\mathcal{D}) \mathrm{d}w \geq 1-\alpha \; . \tag{35}$$

- Consider the observations $\{y_i\}_{i=1}^N$ be i.i.d from

$$Y_i = \mathrm{Ber}(q)\ ,\ \ q \in [0,1] \text{ is the parameter to infer .} \tag{36}$$

- We chose as a prior Beta(α, β), for $\alpha, \beta > 0$:

$$p(q|y) \propto q^{\alpha-1+\sum_{i=1}^N y_i}(1-q)^{\beta-1+N-\sum_{i=1}^N y_i}\ . \tag{37}$$

- The posterior distribution for q is Beta($\alpha + \sum_{i=1}^N y_i, \beta + N - \sum_{i=1}^N y_i$).
- MAP:

$$\hat{w}_{\text{MAP}} = \frac{\alpha - 1 + \sum_i y_i}{\alpha + \beta - 2 + N}\ . \tag{38}$$

- Posterior mean:

$$\hat{w}_{\text{post}} = \frac{\alpha + \sum_i y_i}{\alpha + \beta + N}\ . \tag{39}$$

- We consider the statistical model associated with the LSE is:

$$Y_i \overset{\text{iid}}{\sim} \sum_{i=1}^{d} w_j \phi_j(X_i) + \sigma^2 Z_i \ , \quad i \in \{1, \ldots, N\} \ , \tag{40}$$

  where
  - $Z_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$;
  - X_i are i.i.d random variables with unknown distribution;
  - w is the parameter to infer.
- What does it mean to choose a prior on the hypothesis here?

Prior on parameters induce priors on functions

- A hypothesis f_w is a choice of a model structure $\mathcal{F}$ (first block) and a parameter value (second block) w.
- Consider the linear regression example:

$$f_w(x) = \sum_{i=1}^{d} w_j \phi_j(x) \ , \tag{41}$$

- The number d and the choices of basis functions $\{\phi_j\}$ constitute the model structure;
- The coefficient w, the parameter value.
- Setting a prior $p(w)$ determines what functions this model can generate.
- For the moment $\mathcal{F}$ is fixed but we can also set a prior on the model structure (see after)!
- What is the posterior in this case? ✎

- The likelihood setting $\beta = \sigma^{-2}$ the precision:

$$
\begin{aligned}
\mathrm{L}(\mathcal{D}|w) &= \mathsf{N}_n(\text{vector}(f_w(x_i)), \beta^{-1}\mathrm{I}_n) && (42)\\
&= (\beta/2\pi)^{N/2} \exp\left(-\frac{\beta}{2}\sum_{i=1}^{N}(y_i - \phi(x_i)^{\mathrm{T}} w)^2\right) && (43)\\
&= (\beta/2\pi)^{N/2} \exp\left(-\frac{\beta}{2}\|y - \Phi_x w\|^2\right) \ , && (44)
\end{aligned}
$$

with $\Phi_x =$??.

- If we choose $p(w) = \mathsf{N}_d(m_0, \mathsf{S}_0)$, we get

$$
p(w|\mathcal{D}) = \mathsf{N}(m_N, \mathsf{S}_n) \ , \qquad (45)
$$

Proof of (49). □

- The likelihood setting $\beta = \sigma^{-2}$ the precision:

$$\mathrm{L}(\mathcal{D}|w) = \mathsf{N}_n(\text{vector}(f_w(x_i)), \beta^{-1}\mathrm{I}_n) \tag{46}$$
$$= (\beta/2\pi)^{N/2} \exp\left(-\frac{\beta}{2}\sum_{i=1}^{N}(y_i - \phi(x_i)^{\mathrm{T}} w)^2\right) \tag{47}$$
$$= (\beta/2\pi)^{N/2} \exp\left(-\frac{\beta}{2}\|y - \Phi_x w\|^2\right) , \tag{48}$$

with $\Phi_x =$??.

- If we choose $p(w) = \mathsf{N}_d(m_0, \mathsf{S}_0)$, we get

$$\begin{aligned} &p(w|\mathcal{D}) = \mathsf{N}(m_N, \mathsf{S}_n) , \\ &m_N = \mathsf{S}_N(\mathsf{S}_0^{-1} m_0 + \beta\Phi_x^{\mathrm{T}} y) , \quad \mathsf{S}_N = (\mathsf{S}_0^{-1} + \beta\Phi_x^{\mathrm{T}}\Phi_x)^{-1} . \end{aligned} \tag{49}$$

Proof of (49).

□

Prediction using Bayesian inference

- Consider the linear regression example:

$$f_w(x) = \sum_{i=1}^{d} w_j \phi_j(X_i) \ . \tag{50}$$

- Based on the posterior $p(w|\mathcal{D})$, $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^{N}$, how to make our predictions?
- First frequentist-like option:

$$y_{\mathbf{pred}} = f_{\hat{w}}(x_{\mathbf{new}}) \ , \tag{51}$$

  where $\hat{w}$ is either the MAP or the posterior mean.
- Not really Bayesian...
- Indeed, Bayesian inference is also guided by the aim to give an "optimal" prediction.
- To define what we mean by an "optimal" prediction, we rely on decision theory.

Decision theory for prediction

- Given a dataset $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^N$ with a probabilistic model

$$\{(\tilde{x}, \tilde{y}) \mapsto \mathrm{L}_w(\tilde{x}, \tilde{y}) \,:\, w \in \Theta\} \,,$$

  we would like to find the best estimator for the prediction $y_{\mathbf{pred}}$ based on $x_{\mathbf{new}}$.
- By estimator, here, we mean a function $\mathcal{D} \mapsto \hat{y}^{\mathcal{D}}$ which outputs a function:

$$y_{\mathbf{pred}} = \hat{y}^{\mathcal{D}}(x_{\mathbf{new}}) \,. \tag{52}$$

- How to compare estimator?
- We need a loss function $\ell : \mathsf{Y} \times \mathsf{Y} \to \mathbb{R}_+$ and a prior on w, $w \mapsto p(w)$.
- We define then the conditional risk (given $\mathcal{D}$ and w) as

$$\begin{aligned} \mathrm{cR}(\hat{y}^{\mathcal{D}}, w) &= \mathbb{E}_{(Y_{\mathbf{new}}, X_{\mathbf{new}}) \sim \mathrm{L}_w}[\ell(Y_{new}, \hat{y}^{\mathcal{D}}(X_{\mathbf{new}}))] \qquad (53) \\ &= \int \ell(y_{\mathbf{new}}, \hat{y}^{\mathcal{D}}(x_{\mathbf{new}}))\mathrm{L}_w(x_{\mathbf{new}}, y_{\mathbf{new}})\mathrm{d}(x_{\mathbf{new}}, y_{\mathbf{new}}) \,. \qquad (54) \end{aligned}$$

Decision theory for prediction

- Given a dataset $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^N$ with a probabilistic model

$$\{(\tilde{x}, \tilde{y}) \mapsto \mathrm{L}_w(\tilde{x}, \tilde{y}) \, : \, w \in \Theta\} \, ,$$

  we would like to find the best estimator for the prediction $y_{\mathbf{pred}}$ based on $x_{\mathbf{new}}$.
- By estimator, here, we mean a function $\mathcal{D} \mapsto \hat{y}^{\mathcal{D}}$ which outputs a function:

$$y_{\mathbf{pred}} = \hat{y}^{\mathcal{D}}(x_{\mathbf{new}}) \, . \tag{55}$$

- We define then the conditional risk (given $\mathcal{D}$ and w) as

$$\mathrm{cR}(\hat{y}^{\mathcal{D}}, w) = \mathbb{E}_{(Y_{\mathbf{new}}, X_{\mathbf{new}}) \sim \mathrm{L}_w}[\ell(Y_{new}, \hat{y}^{\mathcal{D}}(X_{\mathbf{new}}))] \, . \tag{56}$$

- An ideal estimator is the one which minimizes the integrated/Bayesian risk:

$$\mathrm{IR} = \mathbb{E}_{\mathcal{D}, w}[\mathrm{R}(\hat{y}^{\mathcal{D}}, w)] = \int \mathrm{R}(\hat{y}^{\mathcal{D}}, w) \mathrm{L}_w(\mathcal{D}) p(w) \mathrm{d}\mathcal{D} \mathrm{d}w \, . \tag{57}$$

- Here $\mathrm{L}_w(\mathcal{D})$ is the complete likelihood $\mathrm{L}_w(x, y) = \prod_{i=1}^N \mathrm{L}_w(x_i, y_i)$.

- Given a dataset $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^N$ with a probabilistic model $\{(\tilde{x}, \tilde{y}) \mapsto \mathrm{L}_w(\tilde{x}, \tilde{y}) \, : \, w \in \Theta\}$, we would like to find the best estimator for the prediction $y_{\mathbf{pred}}$ based on $x_{\mathbf{new}}$.
- By estimator, here, we mean a function $\mathcal{D} \mapsto \hat{y}^{\mathcal{D}}$ which outputs a function:

$$y_{\mathbf{pred}} = \hat{y}^{\mathcal{D}}(x_{\mathbf{new}}) \ . \tag{58}$$

- In the case $\ell(y_1, y_2) = (y_1 - y_2)^2/2$, we can show that the best estimator is

$$\hat{y}^{\mathcal{D}}_{\star,\mathrm{L}^2} = \int \tilde{y}_{\mathbf{new}} \mathrm{L}_w(\tilde{y}_{\mathbf{new}} | x_{\mathbf{new}}) p(w|\mathcal{D}) \mathrm{d}(\tilde{y}_{\mathbf{new}}, w) \ , \tag{59}$$

  where $p(w|\mathcal{D})$ is the posterior distribution associated with prior p.
- It is called the Bayes estimator.

Posterior predictive distribution

- Given a dataset $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^N$ with a probabilistic model $\{(\tilde{x}, \tilde{y}) \mapsto \mathrm{L}_w(\tilde{x}, \tilde{y}) \,:\, w \in \Theta\}$, we would like to find the best estimator for the prediction y_{pred} based on x_{new}.
- By estimator, here, we mean a function $\mathcal{D} \mapsto \hat{y}^{\mathcal{D}}$ which outputs a function:

$$y_{\mathsf{pred}} = \hat{y}^{\mathcal{D}}(x_{\mathsf{new}}) \,. \tag{60}$$

- In the case $\ell(y_1, y_2) = (y_1 - y_2)^2/2$, we can show that the best estimator is

$$\hat{y}^{\mathcal{D}}_{\star,\mathrm{L}^2} = \int \tilde{y}_{\mathsf{new}} \mathrm{L}_w(\tilde{y}_{\mathsf{new}} | x_{\mathsf{new}}) p(w|\mathcal{D}) \mathrm{d}(\tilde{y}_{\mathsf{new}}, w) \,, \tag{61}$$

  where $p(w|\mathcal{D})$ is the posterior distribution associated with prior p.
- This gives rise to the posterior predictive distribution:

$$p_{\mathsf{post}}(\tilde{y}_{\mathsf{new}}|\mathcal{D}) = \int \mathrm{L}_w(\tilde{y}_{\mathsf{new}} | x_{\mathsf{new}}) p(w|\mathcal{D}) \mathrm{d}(w) \,. \tag{62}$$

- With this notation:

$$\hat{y}^{\mathcal{D}}_{\star,\mathrm{L}^2} = \int \tilde{y}_{\mathsf{new}} p_{\mathsf{post}}(\tilde{y}_{\mathsf{new}}|\mathcal{D}, x_{\mathsf{new}}) \mathrm{d}\tilde{y}_{\mathsf{new}} \,. \tag{63}$$

- This distribution give a point estimate for our prediciton but also completely characterizes the uncertainties about our predictions!

Posterior predictive distribution

- Given a dataset $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^N$ with a probabilistic model $\{(\tilde{x}, \tilde{y}) \mapsto \mathrm{L}_w(\tilde{x}, \tilde{y}) \,:\, w \in \Theta\}$, we would like to find the best estimator for the prediction y_{pred} based on x_{new}.
- By estimator, here, we mean a function $\mathcal{D} \mapsto \hat{y}^{\mathcal{D}}$ which outputs a function:

$$y_{\mathsf{pred}} = \hat{y}^{\mathcal{D}}(x_{\mathsf{new}}) \,. \tag{64}$$

- In the case $\ell(y_1, y_2) = (y_1 - y_2)^2/2$, we can show that the best estimator is

$$\hat{y}^{\mathcal{D}}_{\star,\mathrm{L}^2}(x_{\mathsf{new}}) = \int \tilde{y}_{\mathsf{new}} \mathrm{L}_w(\tilde{y}_{\mathsf{new}} | x_{\mathsf{new}}) p(w|\mathcal{D}) \mathrm{d}(\tilde{y}_{\mathsf{new}}, w) \,, \tag{65}$$

where $p(w|\mathcal{D})$ is the posterior distribution associated with prior p.

Proof of (65). $\square$

Predictive posterior for the linear regression model

- The likelihood setting $\beta = \sigma^{-2}$ the precision:

$$\mathrm{L}(\mathcal{D}|w) = \mathsf{N}_n(\text{vector}(f_w(x_i)), \beta^{-1}\mathrm{I}_n) \tag{66}$$

$$= (\beta/2\pi)^{N/2} \exp\left(-\frac{\beta}{2N}\sum_{i=1}^{N}(y_i - f_w(x_i))^2\right) . \tag{67}$$

- If we choose $p(w) = \mathsf{N}_d(m_0, \mathsf{S}_0)$, we get

$$p(w|\mathcal{D}) = \mathsf{N}(m_N, \mathsf{S}_n) , \quad m_N = \mathsf{S}_N(\mathsf{S}_0^{-1}m_0 + \beta\Phi_x^{\mathrm{T}}y) , \quad S_n = (\mathsf{S}_0^{-1} + \beta\Phi_x^{\mathrm{T}}\Phi_x)^{-1} . \tag{68}$$

- Since $p(y_{\mathsf{new}}|w, x_{\mathsf{new}}) = \mathsf{N}(f_w(x_{\mathsf{new}}), \beta^{-1})$, we get that the predictive posterior is

$$p(y_{\mathsf{new}}|x_{\mathsf{new}}) = \mathsf{N}(\phi(x_{\mathsf{new}})^{\mathrm{T}}m_N, \beta^{-1} + \phi(x_{\mathsf{new}})S_N\phi(x_{\mathsf{new}})) , \tag{69}$$

- the Bayes estimator:

$$\hat{y}_{\star,\mathrm{L}^2} = \phi(x_{\mathsf{new}})^{\mathrm{T}}m_N . \tag{70}$$

- Proof in practical sessions.

- What if we are unsure which model is right? So far we assumed we were able to start by making a definite choice of model.
- We can compare models based on marginal likelihoods (also known as model evidence) for each model - this is the probability the model assigns to the observed data.
- This is the normalizing constant in Bayes rule which we ignored previously.

- Let us say that we have two models $\mathcal{F}_1, \mathcal{F}_2$.
- Question: given some data, can we say if one of them is most probable?
- Examples:

$$\mathcal{F}_1 = \left\{ f_w = \sum_{i=1}^{d_1} w_j \phi_j(X_i) \; : \; w \in \mathbb{R}^{d_1} \right\} , \quad \mathcal{F}_2 = \left\{ f_w = \sum_{i=1}^{d_2} w_j \phi_j(X_i) \; : \; w \in \mathbb{R}^{d_2} \right\} \tag{71}$$

  d_1 or d_2 should be privileged?
- Solution: Bayesian paradigm.
- We treat the prior $p(w|\mathcal{F}_1), p(w|\mathcal{F}_2)$ used for $\mathcal{F}_1, \mathcal{F}_2$ as likelihood/conditional probability.
- We set some prior on the models $\mathcal{F}_i$, $i = 1, 2$.

- Let us say that we have two models $\mathcal{F}_1, \mathcal{F}_2$.
- Question: given some data, can we say if one of them is most probable?
- Solution: Bayesian paradigm.
- We treat the likelihoods $p(\mathcal{D}|w, \mathcal{F}_1), p(\mathcal{D}|w, \mathcal{F}_1)$ and priors $p(w|\mathcal{F}_1), p(w|\mathcal{F}_2)$ used for $\mathcal{F}_1, \mathcal{F}_2$ as likelihood/conditional probability.
- We set some prior on the models $\mathcal{F}_i$, $i = 1, 2$.
- In most cases, the uniform prior is chosen...
- The posterior distribution for $(w, \mathcal{F}_i)$ is by Bayes theorem:

$$p(w, \mathcal{F}_i|\mathcal{D}) = p(\mathcal{F}_i)p(w|\mathcal{F}_i)p(\mathcal{D}|w, \mathcal{F}_i)/p(\mathcal{D}) \ , \tag{72}$$

$$p(\mathcal{D}) = \sum_i \int_w p(\mathcal{F}_i)p(w|\mathcal{F}_i)p(\mathcal{D}|w, \mathcal{F}_i)\mathrm{d}w \ . \tag{73}$$

- The posterior distribution for $\mathcal{F}_i$ is then:

$$p(\mathcal{F}_i|\mathcal{D}) = \frac{p(\mathcal{F}_i)}{p(\mathcal{D})} \int p(w|\mathcal{F}_i)p(\mathcal{D}|w, \mathcal{F}_i)\mathrm{d}w \ . \tag{74}$$

- Let us say that we have two models $\mathcal{F}_1, \mathcal{F}_2$.
- Question: given some data, can we say if one of them is most probable?
- The posterior distribution for $(w, \mathcal{F}_i)$ is by Bayes theorem:

$$p(w, \mathcal{F}_i|\mathcal{D}) = p(\mathcal{F}_i)p(w|\mathcal{F}_i)p(\mathcal{D}|w, \mathcal{F}_i)/p(\mathcal{D}) \,, \tag{75}$$

$$p(\mathcal{D}) = \sum_i \int_w p(\mathcal{F}_i)p(w|\mathcal{F}_i)p(\mathcal{D}|w, \mathcal{F}_i)\mathrm{d}w \,. \tag{76}$$

- The posterior distribution for $\mathcal{F}_i$ is then:

$$p(\mathcal{F}_i|\mathcal{D}) = \frac{p(\mathcal{F}_i)\mathrm{Z}_i(\mathcal{D})}{p(\mathcal{D})} \,, \quad \mathrm{Z}_i(\mathcal{D}) = \int p(w|\mathcal{F}_i)p(\mathcal{D}|w, \mathcal{F}_i)\mathrm{d}w \,. \tag{77}$$

- If $p(\mathcal{F}_i) = 1/2$, the posterior distribution for $\mathcal{F}_i$ simplifies:

$$p(\mathcal{F}_i|\mathcal{D}) = \frac{\mathrm{Z}_i(\mathcal{D})}{\mathrm{Z}_1(\mathcal{D}) + \mathrm{Z}_2(\mathcal{D})} \,. \tag{78}$$

- This easily generalizes to finite number of models.

- From looking at the equation of posterior distribution, the marginal likelihood is given by

$$\mathrm{Z}(\mathcal{D}, \mathcal{M}) = \int p(\mathcal{D}|w, \mathcal{M}) p(w|\mathcal{M}) \mathrm{d}w = \int \mathrm{L}_w^{\mathcal{M}}(x, y) p(w|\mathcal{M}) \mathrm{d}w \ . \quad (79)$$

- Second level inference : model comparison

$$p(\mathcal{M}|\mathcal{D}) \propto \mathrm{Z}(\mathcal{D}, \mathcal{M}) p(\mathcal{M}) \ . \quad (80)$$

- Represents some belief/probability on our models given $\mathcal{D}$.
- Model selection:

$$\mathcal{M}^\star = \underset{\mathcal{M}}{\mathrm{argmax}} \, p(\mathcal{M}|\mathcal{D}) \ . \quad (81)$$

Marginal likelihood for the linear regression model

- The likelihood setting $\beta = \sigma^{-2}$ the precision:

$$\mathrm{L}(\mathcal{D}|w) = \mathsf{N}_n(\text{vector}(f_w(x_i)), \beta^{-1}\mathrm{I}_n) \tag{82}$$

$$= (\beta/2\pi)^{N/2} \exp\left(-\frac{\beta}{2N}\sum_{i=1}^{N}(y_i - f_w(x_i))^2\right) . \tag{83}$$

- If we choose $p(w) = \mathsf{N}_d(m_0, \mathrm{S}_0)$, we get

$$p(w|\mathcal{D}) = \mathsf{N}(m_N, \mathrm{S}_n) , \quad m_N = \mathrm{S}_N(\mathrm{S}_0^{-1}m_0 + \beta\Phi_x^{\mathrm{T}}y) , \quad S_n = (\mathrm{S}_0^{-1} + \beta\Phi_x^{\mathrm{T}}\Phi_x)^{-1} . \tag{84}$$

- The marginal likelihood is in the case $p(w) = \mathsf{N}_d(0, \alpha^{-1}\mathrm{I}_d)$:

$$\mathrm{Z}(\mathcal{D}) = \alpha^{d/2}\beta^{N/2}(2\pi)^{-N/2}[\det S_N]^{1/2}\exp\left(-\beta\left\|y\right\|^2/2 + \beta\left\langle S_N m_N, \Phi_x^{\mathrm{T}}y\right\rangle/2\right) \tag{85}$$

$$= \alpha^{d/2}\beta^{N/2}(2\pi)^{-N/2}[\det S_N]^{1/2}\exp\left(-\beta\left\|y - \Phi_x m_N\right\|^2/2 - \alpha\left\|m_N\right\|^2/2\right) . \tag{86}$$

- Proof in small classes.

Hyper-parameter selection

- Let us say that we have some hyperparameters β and α for the likelihoods $p_w((x,y)|\beta)$ and the prior $p(w|\alpha)$ respectively.
- Question: given some data, can we make some recommandations on the choice of these hyperparameters?
- Examples: linear regression (again!) (recall $f_w(x_1) = \sum_{j=1}^{d} w_j \phi_j(x_1)$)

$$p_w((x_1, y_1)|\beta) = (2\pi\sigma^2)^{1/2} \exp(-(y_1 - f_w(x_i))^2/(2\sigma^2)) \ , \ \beta = \sigma^{-2} \ , \quad (87)$$

$$p(w|\alpha) = \alpha \|w\|^2 \ . \quad (88)$$

- Solution: Bayesian paradigm (again...).
- We set some prior on α and β and treat them as parameter as it was the case for models.
- In most cases, the uniform prior is chosen uniform $p(\alpha) = 1$, $p(\beta) = 1$ (even if they do not define a well-defined distribution...).

Hyper-parameter selection

- Let us say that we have some hyperparameters β and α for the likelihoods $p_w((x,y)|\beta)$ and the prior $p(w|\alpha)$ respectively.
- Question: given some data, can we make some recommandations on the choice of these hyperparameters?
- The posterior distribution for (w, λ, β) is by Bayes theorem:

$$p(w, \lambda, \beta|\mathcal{D}) = p(\alpha)p(\beta)p(w|\alpha,\beta)p(\mathcal{D}|w,\alpha,\beta)/p(\mathcal{D}) \ , \tag{89}$$

$$p(\mathcal{D}) = \sum_i \int_w p(\alpha)p(\beta)p(w|\alpha,\beta)p(\mathcal{D}|w,\alpha,\beta)\mathrm{d}w \tag{90}$$

- The posterior distribution for α, β is then:

$$p(\alpha,\beta|\mathcal{D}) = \frac{p(\alpha)p(\beta)}{p(\mathcal{D})} \int p(w|\alpha,\beta)p(\mathcal{D}|w,\alpha,\beta)\mathrm{d}w \ . \tag{91}$$

- Pragamtic choice:

$$(\hat{\alpha}, \hat{\beta}) \in \operatorname{argmax} p(\alpha,\beta|\mathcal{D}) \ , \tag{92}$$

  this corresponds maximization of the marginal likelihood or empirical Bayes approach.
- Example: Bayesian linear regression ✎.

- For most Bayesian inference problems, the integrals needed to do inference and prediction are not analytically tractable - hence the need for variaous approximations.
- Most of the exceptions involve conjugate priors, which combine nicely with the likelihood to give a posterior distribution of the same form.
- Basic Idea : Given likelihood function $L_w(x, y)$, choose a family of prior distributions such that integrals can be obtained tractably.
- If the prior $p(w)$ and posterior $p(w|\mathcal{D})$ belong to same family of distributions, the prior is called a conjugate prior.
- Example: if likelihood function is Gaussian, choosing Gaussian prior over mean will ensure that the posterior distribution is also Gaussian.

Monte Carlo needs: Representing Prior and Posterior by Samples

- The complex distributions we will often use as priors, or obtain as posteriors, may not be easily represented.
- A general technique is to represent a distribution by sampling of many values drawn randomly from it. We can then
  - Visualize the distribution by viewing these sample values, or low dimensional projections of them (PCA..later).
  - Make Monte Carlo estimates for probabilities or expectations with respect to the distribution, by taking averages over these sample values.
- Obtaining a sample from the prior is easy! Obtaining a sample from the posterior is usually more difficult - nevertheless a dominant approach to Bayesian computation.

Kevin P. Murphy. *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press, 2013. ISBN: 9780262018029 0262018020.