Math4AI Lecture No.2

Tin Hadži Veljković
t.hadziveljkovic@uva.nl

20 September 2024.

Table of Contents

Basic Probability Rules

- Probability Notation:
  - $p(x)$: Probability of event x.
  - $p(x, y)$: Joint probability of events x and y.
  - $p(x|y)$: Conditional probability of event x given event y.

Basic Probability Rules

- Probability Notation:
  - $p(x)$: Probability of event x.
  - $p(x, y)$: Joint probability of events x and y.
  - $p(x|y)$: Conditional probability of event x given event y.
- Chain Rule (Two Variables):

$$p(x, y) = p(x|y) \cdot p(y) = p(y|x) \cdot p(x)$$

Basic Probability Rules

- Probability Notation:
  - $p(x)$: Probability of event x.
  - $p(x, y)$: Joint probability of events x and y.
  - $p(x|y)$: Conditional probability of event x given event y.
- Chain Rule (Two Variables):

$$p(x, y) = p(x|y) \cdot p(y) = p(y|x) \cdot p(x)$$

- Chain Rule (Three Variables):

$$p(x, y, z) = p(x|y, z) \cdot p(y, z) = p(x|y, z) \cdot p(y|z) \cdot p(z)$$

Basic Probability Rules

- Probability Notation:
  - $p(x)$: Probability of event x.
  - $p(x, y)$: Joint probability of events x and y.
  - $p(x|y)$: Conditional probability of event x given event y.
- Chain Rule (Two Variables):

$$p(x, y) = p(x|y) \cdot p(y) = p(y|x) \cdot p(x)$$

- Chain Rule (Three Variables):

$$p(x, y, z) = p(x|y, z) \cdot p(y, z) = p(x|y, z) \cdot p(y|z) \cdot p(z)$$

- Marginal Probability (Discrete case):

$$p(x) = \sum_y p(x, y) = \sum_y p(x|y) \cdot p(y)$$

Basic Probability Rules

- Probability Notation:
  - $p(x)$: Probability of event x.
  - $p(x, y)$: Joint probability of events x and y.
  - $p(x|y)$: Conditional probability of event x given event y.
- Chain Rule (Two Variables):

$$p(x, y) = p(x|y) \cdot p(y) = p(y|x) \cdot p(x)$$

- Chain Rule (Three Variables):

$$p(x, y, z) = p(x|y, z) \cdot p(y, z) = p(x|y, z) \cdot p(y|z) \cdot p(z)$$

- Marginal Probability (Discrete case):

$$p(x) = \sum_y p(x, y) = \sum_y p(x|y) \cdot p(y)$$

- Marginal Probability (Continuous case):

$$p(x) = \int p(x, y)\, \mathrm{d}y = \int p(x|y) \cdot p(y)\, \mathrm{d}y$$

Statistical modeling

- *Idea*: Construct a model $\mathcal{M}$ that describes inherent uncertainty/variability in the observed data $\mathcal{D}$, we wish to model the process that generated the data.

Statistical modeling

- *Idea*: Construct a model $\mathcal{M}$ that describes inherent uncertainty/variability in the observed data $\mathcal{D}$, we wish to model the process that generated the data.
- Given a model $\mathcal{M}$ with its parameters $\mathbf{w}$ and data $\mathcal{D}$ we can write the *likelihood function* $p(\mathcal{D}|\mathbf{w}, \mathcal{M})$.

Statistical modeling

- *Idea*: Construct a model $\mathcal{M}$ that describes inherent uncertainty/variability in the observed data $\mathcal{D}$, we wish to model the process that generated the data.
- Given a model $\mathcal{M}$ with its parameters $\mathbf{w}$ and data $\mathcal{D}$ we can write the *likelihood function* $p(\mathcal{D}|\mathbf{w}, \mathcal{M})$.
- *Point estimations*: finding specific optimal parameters $\mathbf{w}^*$ that maximize/minimize some function.

Statistical modeling

- *Idea*: Construct a model $\mathcal{M}$ that describes inherent uncertainty/variability in the observed data $\mathcal{D}$, we wish to model the process that generated the data.
- Given a model $\mathcal{M}$ with its parameters $\mathbf{w}$ and data $\mathcal{D}$ we can write the *likelihood function* $p(\mathcal{D}|\mathbf{w}, \mathcal{M})$.
- *Point estimations*: finding specific optimal parameters $\mathbf{w}^*$ that maximize/minimize some function.
- For example, a *maximum likelihood* solution corresponds to the weights $\mathbf{w}^*$ that maximize the likelihood (or log-likelihood).

Bayesian Inference in Machine Learning

- Objective:
  - Let's assume we have data $\mathcal{D}$ and a model $\mathcal{M}$ with parameters $\mathbf{w}$.
  - The goal is to find the posterior distribution over model parameters $\mathbf{w}$.
  - **Disclaimer:** model parameters $\mathbf{w} \neq$ model $\mathcal{M}$!

Bayesian Inference in Machine Learning

- Objective:
  - Let's assume we have data $\mathcal{D}$ and a model $\mathcal{M}$ with parameters $\mathbf{w}$.
  - The goal is to find the posterior distribution over model parameters $\mathbf{w}$.
  - **Disclaimer:** model parameters $\mathbf{w} \neq$ model $\mathcal{M}$!
- The posterior is given by:

$$p(\mathbf{w}|\mathcal{D}, \mathcal{M}) =$$

Bayesian Inference in Machine Learning

- Objective:
  - Let's assume we have data $\mathcal{D}$ and a model $\mathcal{M}$ with parameters $\mathbf{w}$.
  - The goal is to find the posterior distribution over model parameters $\mathbf{w}$.
  - **Disclaimer:** model parameters $\mathbf{w} \neq$ model $\mathcal{M}$!
- The posterior is given by:

$$p(\mathbf{w}|\mathcal{D}, \mathcal{M}) = \frac{\overbrace{p(\mathcal{D}|\mathbf{w}, \mathcal{M})}^{\text{Likelihood}} \cdot \overbrace{p(\mathbf{w}|\mathcal{M})}^{\text{Prior}}}{\underbrace{p(\mathcal{D}|\mathcal{M})}_{\text{Evidence}}}$$

Bayesian Inference in Machine Learning

- Objective:
  - Let's assume we have data $\mathcal{D}$ and a model $\mathcal{M}$ with parameters $\mathbf{w}$.
  - The goal is to find the posterior distribution over model parameters $\mathbf{w}$.
  - **Disclaimer:** model parameters $\mathbf{w} \neq$ model $\mathcal{M}$!
- The posterior is given by:

$$p(\mathbf{w}|\mathcal{D},\mathcal{M}) = \frac{\overbrace{p(\mathcal{D}|\mathbf{w},\mathcal{M})}^{\text{Likelihood}} \cdot \overbrace{p(\mathbf{w}|\mathcal{M})}^{\text{Prior}}}{\underbrace{p(\mathcal{D}|\mathcal{M})}_{\text{Evidence}}}$$

- The evidence is given by:

$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\mathbf{w},\mathcal{M})p(\mathbf{w}|\mathcal{M})\,\mathrm{d}\mathbf{w}$$

Bayesian Inference in Machine Learning

- Objective:
  - Let's assume we have data $\mathcal{D}$ and a model $\mathcal{M}$ with parameters $\mathbf{w}$.
  - The goal is to find the posterior distribution over model parameters $\mathbf{w}$.
  - **Disclaimer:** model parameters $\mathbf{w} \neq$ model $\mathcal{M}$!
- The posterior is given by:

$$p(\mathbf{w}|\mathcal{D},\mathcal{M}) = \frac{\overbrace{p(\mathcal{D}|\mathbf{w},\mathcal{M})}^{\text{Likelihood}} \cdot \overbrace{p(\mathbf{w}|\mathcal{M})}^{\text{Prior}}}{\underbrace{p(\mathcal{D}|\mathcal{M})}_{\text{Evidence}}}$$

- The evidence is given by:

$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\mathbf{w},\mathcal{M})p(\mathbf{w}|\mathcal{M})\,\mathrm{d}\mathbf{w}$$

- Okay... now what?

MLE and MAP

- Evaluating evidence often isn't tractable.

MLE and MAP

- Evaluating evidence often isn't tractable.
- Let's write the log-posterior:

$$\log p(\mathbf{w}|\mathcal{D}, \mathcal{M}) = \log p(\mathcal{D}|\mathbf{w}, \mathcal{M}) + \log p(\mathbf{w}|\mathcal{M}) - \log p(\mathcal{D}|\mathcal{M})$$

MLE and MAP

- Evaluating evidence often isn't tractable.
- Let's write the log-posterior:

$$\log p(\mathbf{w}|\mathcal{D},\mathcal{M}) = \log p(\mathcal{D}|\mathbf{w},\mathcal{M}) + \log p(\mathbf{w}|\mathcal{M}) - \log p(\mathcal{D}|\mathcal{M})$$

- Finding the *maximum a posteriori* estimation has the form:

$$\frac{\mathrm{d}\log p(\mathbf{w}|\mathcal{D},\mathcal{M})}{\mathrm{d}\mathbf{w}} = 0 \rightarrow \frac{\mathrm{d}\log p(\mathcal{D}|\mathbf{w},\mathcal{M})}{\mathrm{d}\mathbf{w}} + \frac{\mathrm{d}\log p(\mathbf{w}|\mathcal{M})}{\mathrm{d}\mathbf{w}} = 0$$

MLE and MAP

- Evaluating evidence often isn't tractable.
- Let's write the log-posterior:

$$\log p(\mathbf{w}|\mathcal{D}, \mathcal{M}) = \log p(\mathcal{D}|\mathbf{w}, \mathcal{M}) + \log p(\mathbf{w}|\mathcal{M}) - \log p(\mathcal{D}|\mathcal{M})$$

- Finding the *maximum a posteriori* estimation has the form:

$$\frac{\mathrm{d} \log p(\mathbf{w}|\mathcal{D}, \mathcal{M})}{\mathrm{d}\mathbf{w}} = 0 \rightarrow \frac{\mathrm{d} \log p(\mathcal{D}|\mathbf{w}, \mathcal{M})}{\mathrm{d}\mathbf{w}} + \frac{\mathrm{d} \log p(\mathbf{w}|\mathcal{M})}{\mathrm{d}\mathbf{w}} = 0$$

- Similarly to MLE, we would obtain specific optimal parameters $\mathbf{w}^*$ that maximize the posterior.

MLE and MAP

- Evaluating evidence often isn't tractable.
- Let's write the log-posterior:

$$\log p(\mathbf{w}|\mathcal{D},\mathcal{M}) = \log p(\mathcal{D}|\mathbf{w},\mathcal{M}) + \log p(\mathbf{w}|\mathcal{M}) - \log p(\mathcal{D}|\mathcal{M})$$

- Finding the *maximum a posteriori* estimation has the form:

$$\frac{\mathrm{d}\log p(\mathbf{w}|\mathcal{D},\mathcal{M})}{\mathrm{d}\mathbf{w}} = 0 \to \frac{\mathrm{d}\log p(\mathcal{D}|\mathbf{w},\mathcal{M})}{\mathrm{d}\mathbf{w}} + \frac{\mathrm{d}\log p(\mathbf{w}|\mathcal{M})}{\mathrm{d}\mathbf{w}} = 0$$

- Similarly to MLE, we would obtain specific optimal parameters $\mathbf{w}^*$ that maximize the posterior.
- *Insight*: when are MLE and MAP solutions equal?

Bayesian predictive distribution

- Let's assume we managed to obtain the posterior $p(\mathbf{w}|\mathcal{D}, \mathcal{M})$ somehow. What can we do with it?

Bayesian predictive distribution

- Let's assume we managed to obtain the posterior $p(\mathbf{w}|\mathcal{D}, \mathcal{M})$ somehow. What can we do with it?
- Given new datapoint $\mathbf{x}$, the predictive distribution for the observation $\mathbf{t}$ is given by:

$$p(\mathbf{t}|\mathbf{x}, \mathcal{D}, \mathcal{M}) = \int p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \mathcal{M}) p(\mathbf{w}|\mathcal{D}, \mathcal{M}) \, \mathrm{d}\mathbf{w}$$

Bayesian predictive distribution

- Let's assume we managed to obtain the posterior $p(\mathbf{w}|\mathcal{D}, \mathcal{M})$ somehow. What can we do with it?
- Given new datapoint $\mathbf{x}$, the predictive distribution for the observation $\mathbf{t}$ is given by:

$$p(\mathbf{t}|\mathbf{x}, \mathcal{D}, \mathcal{M}) = \int p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \mathcal{M}) p(\mathbf{w}|\mathcal{D}, \mathcal{M}) \, \mathrm{d}\mathbf{w}$$

- What are the benefits of this approach?

Bayesian predictive distribution

- Let's assume we managed to obtain the posterior $p(\mathbf{w}|\mathcal{D}, \mathcal{M})$ somehow. What can we do with it?
- Given new datapoint $\mathbf{x}$, the predictive distribution for the observation $\mathbf{t}$ is given by:

$$p(\mathbf{t}|\mathbf{x}, \mathcal{D}, \mathcal{M}) = \int p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \mathcal{M}) p(\mathbf{w}|\mathcal{D}, \mathcal{M}) \, \mathrm{d}\mathbf{w}$$

- What are the benefits of this approach?
- The predictive distribution in case of a Gaussian distributions is given by:

$$p(\mathbf{t}|\mathbf{x}, \mathcal{D}, \mathcal{M}) = \mathcal{N}(\mathbf{t}|\mathbf{m}^T \phi(\mathbf{x}), \sigma_N^2(\mathbf{x})).$$

Bayesian predictive distribution

- Let's assume we managed to obtain the posterior $p(\mathbf{w}|\mathcal{D}, \mathcal{M})$ somehow. What can we do with it?
- Given new datapoint $\mathbf{x}$, the predictive distribution for the observation $\mathbf{t}$ is given by:

$$p(\mathbf{t}|\mathbf{x}, \mathcal{D}, \mathcal{M}) = \int p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \mathcal{M}) p(\mathbf{w}|\mathcal{D}, \mathcal{M}) \, \mathrm{d}\mathbf{w}$$

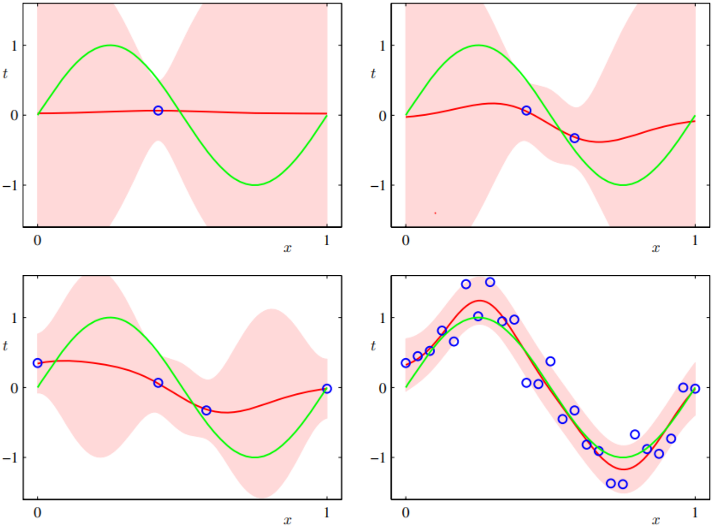
- What are the benefits of this approach?
- The predictive distribution in case of a Gaussian distributions is given by:

$$p(\mathbf{t}|\mathbf{x}, \mathcal{D}, \mathcal{M}) = \mathcal{N}(\mathbf{t}|\mathbf{m}^T \phi(\mathbf{x}), \sigma_N^2(\mathbf{x})).$$

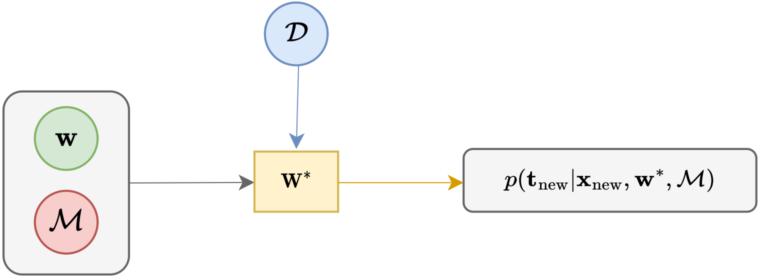
- The variance is a function of datapoints $\mathbf{x}$:

$$\sigma_N^2(\mathbf{x}) = 1/\beta + \phi(\mathbf{x})^\mathsf{T} \mathbf{S}_N \phi(\mathbf{x}).$$

Visualizing uncertainty

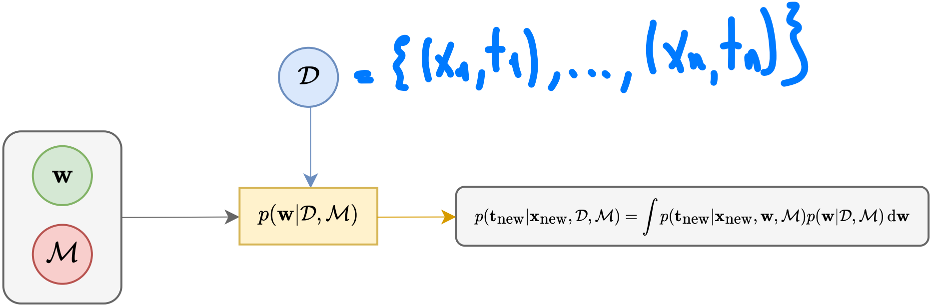


New predictions - point estimations



- Predictions using the likelihood $\mathcal{L}$ with optimal weights $\mathbf{w}^*$.
- Selecting one specific instantiation of the model $\mathcal{M}$, no uncertainty in the model parameters.

New predictions - Bayesian predictive distribution

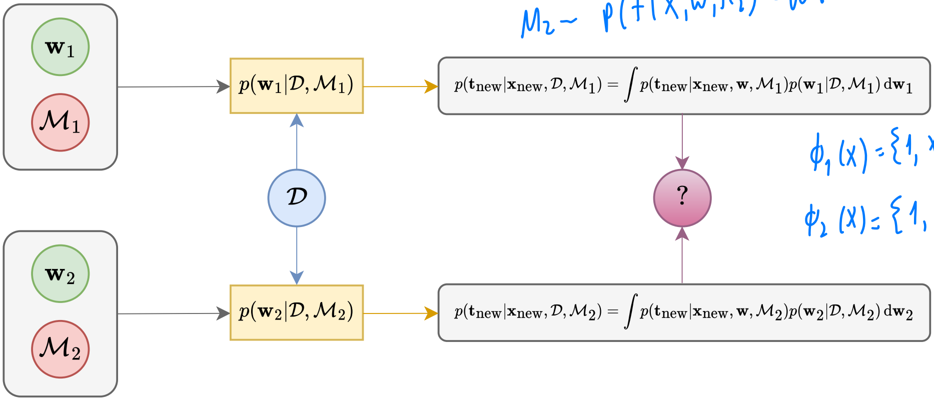


- Obtaining the posterior distribution over model parameters **w**.
- Predictions using the predictive distribution, without choosing a specific set of weights **w**.
- The "average" prediction over all possible model instantiations.

New predictions - combining models?

$\mathcal{M}_1 \sim p(t|x,w,\mathcal{M}_1) = \mathcal{N}(t|w^T\phi_1(x), 1/\beta)$

$\mathcal{M}_2 \sim p(t|x,w,\mathcal{M}_2) = \mathcal{N}(t|w^T\phi_2(x), 1/\gamma)$

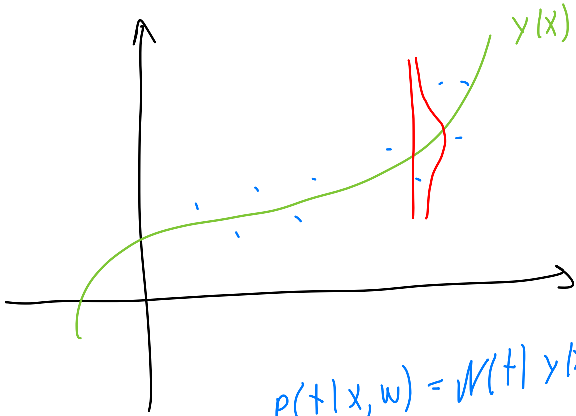


$\phi_1(x) = \{1, x^2, x^4, x^6\}$

$\phi_2(x) = \{1, x^3, x^5, x^7\}$

- Given the same data $\mathcal{D}$, can we combine different "worlds" (models)?

$$y(x) = \phi^T(x)\, w$$



$$p(t \mid x, w) = \mathcal{N}(t \mid y(x), 1/\beta)$$

Back to the evidence!

- The evidence was given by $p(\mathcal{D}|\mathcal{M})$.

Back to the evidence!

- The evidence was given by $p(\mathcal{D}|\mathcal{M})$.
- More generally, for a model $\mathcal{M}_i$, the evidence is $p(\mathcal{D}|\mathcal{M}_i)$.

Back to the evidence!

- The evidence was given by $p(\mathcal{D}|\mathcal{M})$.
- More generally, for a model $\mathcal{M}_i$, the evidence is $p(\mathcal{D}|\mathcal{M}_i)$.
- But what does it mean?

Back to the evidence!

- The evidence was given by $p(\mathcal{D}|\mathcal{M})$.
- More generally, for a model $\mathcal{M}_i$, the evidence is $p(\mathcal{D}|\mathcal{M}_i)$.
- But what does it mean?
- Let's observe the following:

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)\,\mathrm{d}\mathbf{w}$$

Back to the evidence!

- The evidence was given by $p(\mathcal{D}|\mathcal{M})$.
- More generally, for a model $\mathcal{M}_i$, the evidence is $p(\mathcal{D}|\mathcal{M}_i)$.
- But what does it mean?
- Let's observe the following:

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)\,\mathrm{d}\mathbf{w}$$

- The evidence is the measure of how well the model, as a whole, predicts the data.

Back to the evidence!

- The evidence was given by $p(\mathcal{D}|\mathcal{M})$.
- More generally, for a model $\mathcal{M}_i$, the evidence is $p(\mathcal{D}|\mathcal{M}_i)$.
- But what does it mean?
- Let's observe the following:

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)\,\mathrm{d}\mathbf{w}$$

- The evidence is the measure of how well the model, as a whole, predicts the data.
- We can optimize the parameters of a specific model, but can we optimize which **model** to choose?

Bayes factor

- Assume we are given data $\mathcal{D}$ and two models $\mathcal{M}_1$, $\mathcal{M}_2$.

Bayes factor

- Assume we are given data $\mathcal{D}$ and two models $\mathcal{M}_1$, $\mathcal{M}_2$.
- We calculate the evidence of models $\mathcal{M}_1$ and $\mathcal{M}_2$.

Bayes factor

- Assume we are given data $\mathcal{D}$ and two models $\mathcal{M}_1$, $\mathcal{M}_2$.
- We calculate the evidence of models $\mathcal{M}_1$ and $\mathcal{M}_2$.
- Their ratio K is called the *Bayes factor*:

$$K = \frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)}$$

Bayes factor

- Assume we are given data $\mathcal{D}$ and two models $\mathcal{M}_1$, $\mathcal{M}_2$.
- We calculate the evidence of models $\mathcal{M}_1$ and $\mathcal{M}_2$.
- Their ratio K is called the *Bayes factor*:

$$K = \frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)}$$

- Robust, penalizes overfitting. How do we interpret the values?

K Range	Strength of Evidence
1 to 3.2	Not worth more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
> 100	Decisive

Table: Strength of Evidence vs. K

Mixture distributions

- Let's think again of combining models!

Mixture distributions

- Let's think again of combining models!
- Assume we have a set of models $\{\mathcal{M}_i\}$, where $i = 1, \ldots, L$ and data $\mathcal{D}$.

Mixture distributions

- Let's think again of combining models!
- Assume we have a set of models $\{\mathcal{M}_i\}$, where $i = 1, \ldots, L$ and data $\mathcal{D}$.
- Let's find the *model posterior* distribution:

$$p(\mathcal{M}_i|\mathcal{D}) = \frac{\overbrace{p(\mathcal{D}|\mathcal{M}_i)}^{\text{Model evidence}} \cdot \overbrace{p(\mathcal{M}_i)}^{\text{Model prior}}}{\underbrace{p(\mathcal{D})}_{\text{Probability of the data}}}$$

Mixture distributions

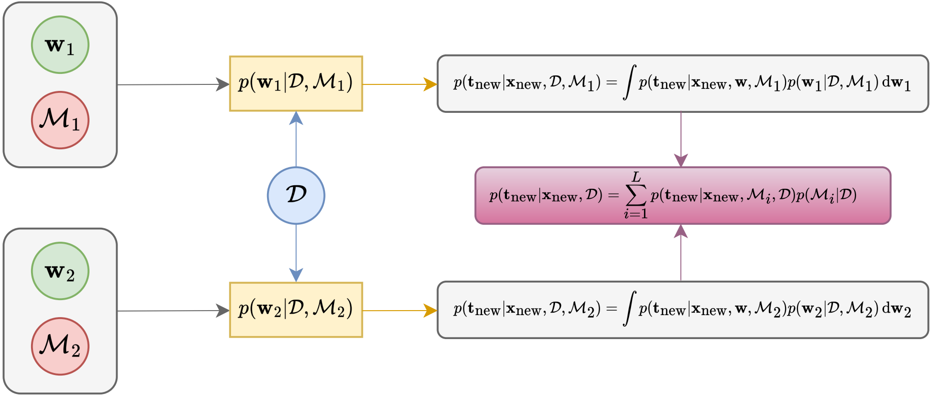
- Let's think again of combining models!
- Assume we have a set of models $\{\mathcal{M}_i\}$, where $i = 1, \ldots, L$ and data $\mathcal{D}$.
- Let's find the *model posterior* distribution:

$$p(\mathcal{M}_i|\mathcal{D}) = \frac{\overbrace{p(\mathcal{D}|\mathcal{M}_i)}^{\text{Model evidence}} \cdot \overbrace{p(\mathcal{M}_i)}^{\text{Model prior}}}{\underbrace{p(\mathcal{D})}_{\text{Probability of the data}}}$$

- A *mixture distribution* that uses all models is given by:

$$p(\mathbf{t}_{\text{new}}|\mathbf{x}_{\text{new}}, \mathcal{D}) = \sum_{i=1}^{L} p(\mathbf{t}_{\text{new}}|\mathbf{x}_{\text{new}}, \mathcal{M}_i, \mathcal{D}) p(\mathcal{M}_i|\mathcal{D}).$$

New predictions - combining models



- New predictions by averaging predictive distributions of individual models, weighted by their posterior probabilities.

Kronecker delta

- Definition of the Kronecker delta:

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

Kronecker delta

- Definition of the Kronecker delta:

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

- Let $\mathbf{a} = (a_1, \ldots, a_n)^T$. Summing over the elements of this vector with δ_{ij} yields:

$$\sum_{i=1}^{n} \delta_{ik} a_i = a_k$$

Kronecker delta

- Similarly, if we have a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with elements A_{ij}, and we sum over all elements of the matrix using one Kronecker delta, we get:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} \delta_{ik} = \sum_{j=1}^{m} A_{kj}$$

Kronecker delta

- Similarly, if we have a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with elements A_{ij}, and we sum over all elements of the matrix using one Kronecker delta, we get:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} \delta_{ik} = \sum_{j=1}^{m} A_{kj}$$

- Sometimes we will encounter summations that include multiple Kronecker delta symbols (both which are include indices of the matrix). In this case, we would get:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} \delta_{ik} \delta_{jl} = \sum_{j=1}^{m} A_{kj} \delta_{jl} = A_{kl}$$

Kronecker delta

- Similarly, if we have a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with elements A_{ij}, and we sum over all elements of the matrix using one Kronecker delta, we get:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} \delta_{ik} = \sum_{j=1}^{m} A_{kj}$$

- Sometimes we will encounter summations that include multiple Kronecker delta symbols (both which are include indices of the matrix). In this case, we would get:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} \delta_{ik} \delta_{jl} = \sum_{j=1}^{m} A_{kj} \delta_{jl} = A_{kl}$$

- Kronecker delta acts as an index selector and eliminates sums over indices which are also part of the Kronecker delta symbols.

Linear Algebra: General notes

- Matrix-vector multiplication: $\mathbf{b} = \mathbf{Xa}$. The i-th element b_i is given by:

$$b_i = \sum_k \mathrm{X}_{ik} a_k$$

Linear Algebra: General notes

- Matrix-vector multiplication: $\mathbf{b} = \mathbf{X}\mathbf{a}$. The i-th element b_i is given by:

$$b_i = \sum_k \mathrm{X}_{ik} a_k$$

- Matrix-matrix multiplication: $\mathbf{C} = \mathbf{A}\mathbf{B}$. The ij-th element C_{ij} is given by:

$$\mathrm{C}_{ij} = \sum_k \mathrm{A}_{ik} \mathrm{B}_{kj}$$

Linear Algebra: General notes

- Matrix-vector multiplication: $\mathbf{b} = \mathbf{X}\mathbf{a}$. The i-th element b_i is given by:

$$b_i = \sum_k \mathrm{X}_{ik} a_k$$

- Matrix-matrix multiplication: $\mathbf{C} = \mathbf{A}\mathbf{B}$. The ij-th element C_{ij} is given by:

$$\mathrm{C}_{ij} = \sum_k \mathrm{A}_{ik} \mathrm{B}_{kj}$$

- Dot product: $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b}$. The resulting scalar is given by:

$$\mathbf{a}^T \mathbf{b} = \sum_k a_k b_k$$

Vector Calculus: General notes

- Suppose we have a vector valued function $\mathbf{f}(\mathbf{X}) = \mathbf{Y}$, where $\mathbf{X} \in \mathbb{R}^n$, and $\mathbf{Y} \in \mathbb{R}^m$. We can write this mapping as $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$

Vector Calculus: General notes

- Suppose we have a vector valued function $\mathbf{f}(\mathbf{X}) = \mathbf{Y}$, where $\mathbf{X} \in \mathbb{R}^n$, and $\mathbf{Y} \in \mathbb{R}^m$. We can write this mapping as $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$
- The derivative $\frac{d\mathbf{f}}{d\mathbf{x}}$ will have, by definition, shape $\mathbb{R}^{m \times n}$.

Vector Calculus: General notes

- Suppose we have a vector valued function $\mathbf{f}(\mathbf{X}) = \mathbf{Y}$, where $\mathbf{X} \in \mathbb{R}^n$, and $\mathbf{Y} \in \mathbb{R}^m$. We can write this mapping as $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$
- The derivative $\frac{d\mathbf{f}}{d\mathbf{x}}$ will have, by definition, shape $\mathbb{R}^{m \times n}$.
- For example, if the function $f \in \mathbb{R}$ (a scalar), and the input is a vector $\mathbf{x} \in R^n$, then the derivative will have the shape $\mathbb{R}^{1 \times n}$ (i.e. the derivative is an n-dimensional row vector).

Vector Calculus: An example from A to Z

- Suppose we have a function $f(\mathbf{X}) = \mathbf{b}^T \mathbf{X}^\mathsf{T} \mathbf{X} \mathbf{c}$

Vector Calculus: An example from A to Z

- Suppose we have a function $f(\mathbf{X}) = \mathbf{b}^T \mathbf{X}^{\mathsf{T}} \mathbf{X} \mathbf{c}$
- We have $f \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^{n \times m}$, while $\mathbf{b} \in \mathbb{R}^m$, and $\mathbf{c} \in \mathbb{R}^m$

Vector Calculus: An example from A to Z

$A \in \mathbb{R}^{n \times m}$

$B \in \mathbb{R}^{m \times k}$

$C = AB \in \mathbb{R}^{n \times k}$

$(n \times m) \cdot (m \times k) = n \times k$

- Suppose we have a function $f(\mathbf{X}) = \mathbf{b}^T \mathbf{X}^\mathsf{T} \mathbf{X} \mathbf{c}$
- We have $f \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^{n \times m}$, while $\mathbf{b} \in \mathbb{R}^m$, and $\mathbf{c} \in \mathbb{R}^m$
- What is the shape of the derivative $\partial f(\mathbf{X}) / \partial \mathbf{X}$?

Vector Calculus: An example from A to Z

- Let's calculate $\partial f(\mathbf{X})/\partial\mathbf{X}$.

Vector Calculus: An example from A to Z

- Let's calculate $\partial f(\mathbf{X})/\partial \mathbf{X}$.
- Before we continue, let's expand the expression for the function $f = \mathbf{b}^T \mathbf{X}^\mathsf{T} \mathbf{X} \mathbf{c}$.

Vector Calculus: An example from A to Z

$b^T X^T v$

- Let's calculate $\partial f(\mathbf{X})/\partial \mathbf{X}$.
- Before we continue, let's expand the expression for the function $f = \mathbf{b}^T \mathbf{X}^\mathsf{T} \mathbf{X} \mathbf{c}$.
- First we will expand the term $\mathbf{Xc}$, which is a matrix-vector multiplication. We will call this new vector $\mathbf{v} = \mathbf{Xc}$. The l-th element of this vector is given by:

$$v_l = \sum_k \mathrm{X}_{lk} c_k$$

Vector Calculus: An example from A to Z

- Let's calculate $\partial f(\mathbf{X})/\partial \mathbf{X}$.
- Before we continue, let's expand the expression for the function $f = \mathbf{b}^T \mathbf{X}^\mathsf{T} \mathbf{X} \mathbf{c}$.
- First we will expand the term $\mathbf{Xc}$, which is a matrix-vector multiplication. We will call this new vector $\mathbf{v} = \mathbf{Xc}$. The l-th element of this vector is given by:

$$v_l = \sum_k \mathrm{X}_{lk} c_k$$

- Next, we see that we can write out the function f as $\mathbf{b}^T \mathbf{X}^\mathsf{T} \mathbf{v}$.

Vector Calculus: An example from A to Z

- Let's calculate $\partial f(\mathbf{X})/\partial \mathbf{X}$.
- Before we continue, let's expand the expression for the function $f = \mathbf{b}^T \mathbf{X}^\mathrm{T} \mathbf{X} \mathbf{c}$.
- First we will expand the term $\mathbf{Xc}$, which is a matrix-vector multiplication. We will call this new vector $\mathbf{v} = \mathbf{Xc}$. The l-th element of this vector is given by:

$$v_l = \sum_k \mathrm{X}_{lk} c_k$$

- Next, we see that we can write out the function f as $\mathbf{b}^T \mathbf{X}^\mathrm{T} \mathbf{v}$.
- Using similar logic, we will write the matrix-vector multiplication $\mathbf{X}^\mathrm{T} \mathbf{v}$ as $\mathbf{w}$. Now, the m-th element of this vector is given by:

$$w_m = \sum_l \mathrm{X}^\mathrm{T}_{ml} v_l = \sum_l \mathrm{X}_{lm} v_l = \sum_l \sum_k \mathrm{X}_{lm} \mathrm{X}_{lk} c_k$$

Vector Calculus: An example from A to Z

- Using this substitution, we can see that the function f can be written as $\mathbf{b}^{\mathsf{T}}\mathbf{w}$, which is just a dot product between the two vectors. Thus, we can write:

$$f = \mathbf{b}^{\mathsf{T}}\mathbf{w} = \sum_{m} b_m w_m = \sum_{l} \sum_{k} \sum_{m} b_m \mathrm{X}_{lm} \mathrm{X}_{lk} c_k$$

Vector Calculus: An example from A to Z

- Using this substitution, we can see that the function f can be written as $\mathbf{b}^\mathsf{T}\mathbf{w}$, which is just a dot product between the two vectors. Thus, we can write:

$$f = \mathbf{b}^\mathsf{T}\mathbf{w} = \sum_m b_m w_m = \sum_l \sum_k \sum_m b_m \mathrm{X}_{lm} \mathrm{X}_{lk} c_k$$

- The derivative will be a matrix, and we find the ij-th element of the derivative by taking the derivative wrt. ij-th element of the matrix $\mathbf{X}$:

$$\left[\frac{\mathrm{d}f}{\mathrm{d}\mathbf{X}}\right]_{ij} = \frac{\partial f}{\partial \mathrm{X}_{ij}}$$

Vector Calculus: An example from A to Z

- Using this substitution, we can see that the function f can be written as $\mathbf{b}^\top \mathbf{w}$, which is just a dot product between the two vectors. Thus, we can write:

$$f = \mathbf{b}^\top \mathbf{w} = \sum_m b_m w_m = \sum_l \sum_k \sum_m b_m \mathrm{X}_{lm} \mathrm{X}_{lk} c_k$$

- The derivative will be a matrix, and we find the ij-th element of the derivative by taking the derivative wrt. ij-th element of the matrix $\mathbf{X}$:

$$\left[\frac{\mathrm{d}f}{\mathrm{d}\mathbf{X}}\right]_{ij} = \frac{\partial f}{\partial \mathrm{X}_{ij}}$$

- Let's take the derivative!

Vector Calculus: An example from A to Z

- The derivative is given by:
- $$\frac{\partial f}{\partial X_{ij}} = \frac{\partial}{\partial X_{ij}} \left(\sum_l \sum_k \sum_m b_m X_{lm} X_{lk} c_k \right)$$

Vector Calculus: An example from A to Z

- The derivative is given by:
- $$\frac{\partial f}{\partial X_{ij}} = \frac{\partial}{\partial X_{ij}} \left(\sum_l \sum_k \sum_m b_m X_{lm} X_{lk} c_k \right)$$
$$= \sum_l \sum_k \sum_m b_m \frac{\partial X_{lm}}{\partial X_{ij}} X_{lk} c_k + \sum_l \sum_k \sum_m b_m X_{lm} \frac{\partial X_{lk}}{\partial X_{ij}} c_k$$

Vector Calculus: An example from A to Z

- The derivative is given by:
-

$$\frac{\partial f}{\partial X_{ij}} = \frac{\partial}{\partial X_{ij}}\left(\sum_l \sum_k \sum_m b_m X_{lm} X_{lk} c_k\right)$$

$$= \sum_l \sum_k \sum_m b_m \frac{\partial X_{lm}}{\partial X_{ij}} X_{lk} c_k + \sum_l \sum_k \sum_m b_m X_{lm} \frac{\partial X_{lk}}{\partial X_{ij}} c_k$$

$$= \sum_l \sum_k \sum_m b_m \delta_{li} \delta_{mj} X_{lk} c_k + \sum_l \sum_k \sum_m b_m X_{lm} \delta_{li} \delta_{kj} c_k$$

Vector Calculus: An example from A to Z

- The derivative is given by:
- $$\frac{\partial f}{\partial X_{ij}} = \frac{\partial}{\partial X_{ij}} \left(\sum_l \sum_k \sum_m b_m X_{lm} X_{lk} c_k \right)$$
$$= \sum_l \sum_k \sum_m b_m \frac{\partial X_{lm}}{\partial X_{ij}} X_{lk} c_k + \sum_l \sum_k \sum_m b_m X_{lm} \frac{\partial X_{lk}}{\partial X_{ij}} c_k$$
$$= \sum_l \sum_k \sum_m b_m \delta_{li} \delta_{mj} X_{lk} c_k + \sum_l \sum_k \sum_m b_m X_{lm} \delta_{li} \delta_{kj} c_k$$
$$= \sum_k b_j X_{ik} c_k + \sum_m X_{im} b_m c_j$$

Vector Calculus: An example from A to Z

- The derivative is given by:
- $$\begin{aligned}
\frac{\partial f}{\partial X_{ij}} &= \frac{\partial}{\partial X_{ij}} \left(\sum_l \sum_k \sum_m b_m X_{lm} X_{lk} c_k \right) \\
&= \sum_l \sum_k \sum_m b_m \frac{\partial X_{lm}}{\partial X_{ij}} X_{lk} c_k + \sum_l \sum_k \sum_m b_m X_{lm} \frac{\partial X_{lk}}{\partial X_{ij}} c_k \\
&= \sum_l \sum_k \sum_m b_m \delta_{li} \delta_{mj} X_{lk} c_k + \sum_l \sum_k \sum_m b_m X_{lm} \delta_{li} \delta_{kj} c_k \\
&= \sum_k b_j X_{ik} c_k + \sum_m X_{im} b_m c_j \\
&= \left(\sum_k X_{ik} c_k \right) b_j + \left(\sum_m X_{im} b_m \right) c_j
\end{aligned}$$

Vector Calculus: An example from A to Z

- Let's define new vectors $\tilde{\mathbf{c}} = \mathbf{Xc}$ and $\tilde{\mathbf{b}} = \mathbf{Xb}$.

Vector Calculus: An example from A to Z

- Let's define new vectors $\tilde{\mathbf{c}} = \mathbf{X}\mathbf{c}$ and $\tilde{\mathbf{b}} = \mathbf{X}\mathbf{b}$.
- Then, their p-th element is given by $\tilde{b}_p = \sum_r \mathrm{X}_{pr} b_r$ and $\tilde{c}_p = \sum_r \mathrm{X}_{pr} c_r$ respectively.

Vector Calculus: An example from A to Z

- Let's define new vectors $\tilde{\mathbf{c}} = \mathbf{X}\mathbf{c}$ and $\tilde{\mathbf{b}} = \mathbf{X}\mathbf{b}$.
- Then, their p-th element is given by $\tilde{b}_p = \sum_r \mathrm{X}_{pr} b_r$ and $\tilde{c}_p = \sum_r \mathrm{X}_{pr} c_r$ respectively.
- We can recognize that in the previous expression!

$$\frac{\partial f}{\partial \mathrm{X}_{ij}} = \left(\sum_k \mathrm{X}_{ik} c_k \right) b_j + \left(\sum_m \mathrm{X}_{im} b_m \right) c_j = \tilde{c}_i b_j + \tilde{b}_i c_j$$

Vector Calculus: An example from A to Z

- Let's define new vectors $\tilde{\mathbf{c}} = \mathbf{X}\mathbf{c}$ and $\tilde{\mathbf{b}} = \mathbf{X}\mathbf{b}$.
- Then, their p-th element is given by $\tilde{b}_p = \sum_r \mathrm{X}_{pr} b_r$ and $\tilde{c}_p = \sum_r \mathrm{X}_{pr} c_r$ respectively.
- We can recognize that in the previous expression!

$$\frac{\partial f}{\partial \mathrm{X}_{ij}} = \left(\sum_k \mathrm{X}_{ik} c_k \right) b_j + \left(\sum_m \mathrm{X}_{im} b_m \right) c_j = \tilde{c}_i b_j + \tilde{b}_i c_j$$

- This is just the outer product, so we have:

$$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{X}} = \tilde{\mathbf{c}}\mathbf{b}^\top + \tilde{\mathbf{b}}\mathbf{c}^\top = \mathbf{X}\mathbf{c}\mathbf{b}^\top + \mathbf{X}\mathbf{b}\mathbf{c}^\top$$

Vector Calculus: An example from A to Z

- Let's define new vectors $\tilde{\mathbf{c}} = \mathbf{X}\mathbf{c}$ and $\tilde{\mathbf{b}} = \mathbf{X}\mathbf{b}$.
- Then, their p-th element is given by $\tilde{b}_p = \sum_r X_{pr} b_r$ and $\tilde{c}_p = \sum_r X_{pr} c_r$ respectively.
- We can recognize that in the previous expression!

$$D_{ij} = \frac{\partial f}{\partial X_{ij}} = \left(\sum_k X_{ik} c_k\right) b_j + \left(\sum_m X_{im} b_m\right) c_j = \tilde{c}_i b_j + \tilde{b}_i c_j$$

- This is just the outer product, so we have:

$$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{X}} = \tilde{\mathbf{c}}\mathbf{b}^{\mathrm{T}} + \tilde{\mathbf{b}}\mathbf{c}^{\mathrm{T}} = \mathbf{X}\mathbf{c}\mathbf{b}^{\mathrm{T}} + \mathbf{X}\mathbf{b}\mathbf{c}^{\mathrm{T}}$$

- Therefore the full derivative is given by:

$$\boxed{\frac{\mathrm{d}f}{\mathrm{d}\mathbf{X}} = \underbrace{\mathbf{X}}_{\in\mathbb{R}^{n\times m}}\underbrace{\left(\mathbf{c}\mathbf{b}^{\mathrm{T}} + \mathbf{b}\mathbf{c}^{\mathrm{T}}\right)}_{\in\mathbb{R}^{m\times m}}} \in \mathbb{R}^{n\times m}$$

$f \in \mathbb{R}$

$X \in \mathbb{R}^{n\times m}$

$\frac{df}{dX} \in \mathbb{R}^{1\times(n\times m)}$

$\mathbb{R}^{n\times m}$

Questions?

Ask me anything!