Introduction to Statistical Learning

CSCC11 – Topic 02

Winter 2025

Computer & Mathematical Sciences
UNIVERSITY OF TORONTO
SCARBOROUGH

Topics

- Basic Probability Rules
  - Basic rules of probability
  - Bayes rule and its variations
- Estimation of Model Parameters
  - MAP
  - ML
  - Bayesian Method
- Probabilistic Interpretation of Linear Regression Model

Probability Rules

Why Probability?

- Probability is for quantifying and reasoning about uncertainty.
- Uncertainties in ML
  - Aleatoric uncertainty / Data uncertainty
    - Intrinsic (irreducible) stochasticity in the data generation process
    - Measurements are noisy.
    - Data distributions overlap
    - Can be improved by better modelling, but not by getting more training data.
  - Epistemic uncertainty / Model uncertainty
    - Uncertainty due to insufficient knowledge of the mapping from input to output
    - Limited and biased data, missing data (gap in the knowledge)
    - Model structure uncertainty
    - Model parameter uncertainty
    - Can be improved by changing the model or getting more training data

Basic Rules

Basic Rules	Probabilities	PDFs
Axioms/Definitions	$P(A) \in [0,1]$	$\forall x,\ p(x) \geq 0,\ P(a < x \leq b) = \int_a^b p(x)dx$
Sum Rule	$P(A) + P(\bar{A}) = 1$	$\int_{-\infty}^{\infty} p(x)dx = 1$
Product Rule	$P(A,B) = P(A\|B)P(B)$	$p(x,y) = p(x\|y)p(y) = p(y\|x)p(x)$
Independence	$P(A,B) = P(A)P(B)$	$p(x,y) = p(x)p(y)$
Marginalization	$P(B) = \sum_i P(A_i B)$	$p(y) = \int_{-\infty}^{\infty} p(x,y)dx$
Conditional on additional information	$P(A,B \mid C) = P(A\|B,C)P(B\|C)$	$p(y\|z) = \int_{-\infty}^{\infty} p(x,y\|z)dx$

Note: In statistics, $p_X(\cdot)$ is for PMF and $f_X(\cdot)$ is for PDF. In ML textbooks ([CB][KM]) and course notes, $p_X(\cdot)$ is used for both PMF and PDF. And we drop the subscript X when the context is clear.

Example

- We roll 2 dice and assume two rolls are independent
- What is the probability that die 1 shows 5?
- What is the probability that die 2 shows 1?
- What is the probability that the sum of the two rolls is 9?
- What is the probability that the sum of the two rolls is 9 given that the die 2 shows 6?

Bayes Rule

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

$$= \frac{P(B|A)P(A)}{\sum_i P(A_i, B)}$$

For Events

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)}$$

$$= \frac{p_{Y|X}(y|x)p_X(x)}{\int_{-\infty}^{+\infty} p(x,y)\,dx}$$

For Probability Density Function

Bayes Rule Variation

$$p_\Theta(\theta) \equiv f_\Theta(\theta) = \lim_{\delta \to 0} \frac{P(\theta \le \Theta \le \theta + \delta)}{\delta}$$

$$p_K(k) = P(K = k)$$

- One Discrete and one continuous random variable
- K: discrete Θ: continuous

$$P(K = k, \theta \le \Theta \le \theta + \delta)$$

$$= P(K = k) P(\theta \le \Theta \le \theta + \delta | K = k) \approx p_K(k) p_{\Theta|K}(\theta|k) \delta$$

$$= P(\theta \le \Theta \le \theta + \delta) P(K = k | \theta \le \Theta \le \theta + \delta) \approx p_\Theta(\theta) p_{K|\Theta}(k|\theta) \delta$$

$$p_{\Theta|K}(\theta|k) = \frac{p_\Theta(\theta) p_{K|\Theta}(k|\theta)}{p_K(k)}$$

$$p_{K|\Theta}(k|\theta) = \frac{p_K(k) p_{\Theta|K}(\theta|k)}{p_\Theta(\theta)}$$

Note: In statistics, $p_X(\cdot)$ is for PMF and $f_X(\cdot)$ is for PDF. In ML textbooks ([CB][KM]) and course notes, $p_X(\cdot)$ is used for both PMF and PDF. And we drop the subscript X when the context is clear.

Bayes Rule Variation – Cont'd

- When the meaning of the random variables are clear from the context, we drop the subscripts in the name of the PMF/PDF
- k is discrete and θ is continuous.

$$p(\theta|k) = \frac{p(\theta)p(k|\theta)}{p(k)}$$

$$p(k|\theta) = \frac{p(k)p(\theta|k)}{p(\theta)}$$

Bayes Rule

Posterior Likelihood Prior

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

evidence

$$p(\mathcal{M}|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M})p(\mathcal{M})$$

- The posterior distribution over the model parameters is the probability after we see the data.
- The prior distribution over the model parameter is our belief of the model parameter before we see the data.
- The likelihood measures the consistency of the data with the model. It is a function of the model, not a function of the data.
- The evidence is the marginal likelihood of data over all possible models. It is a normalizing constant. It does not depend on $\mathcal{M}$.

Bayes Rule

$$p(\mathcal{M}|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M})p(\mathcal{M})$$

Posterior $\propto$ Likelihood $\times$ Prior

To obtain posterior, we need

1. Data: to form likelihood $p(\mathcal{D}|\mathcal{M})$
2. A prior probability distribution, $p(\mathcal{M})$, to represent our prior belief about the model.

Parameter Estimation

Inference

- Statistics:
  - The process of quantifying uncertainty about an unknown quantity estimated from a finite sample of data.
- ML
  - Computing the conditional distribution $p(\mathbf{y} \mid \mathbf{x};\ \boldsymbol{\theta})$
  - Predicting the outcome of unseen data using the trained model

Estimation

- Estimation: finding the best model parameters (a point estimate).
  - MAP (Maximum A Posteriori)

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|\mathcal{D}) = \underset{\theta}{\operatorname{argmax}}(p(\mathcal{D}|\theta)p(\theta))$$

  - ML (Maximum Likelihood)

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D}|\theta)$$

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

If the prior is very uninformative, that is we might believe that the prior is uniformly distributed, then $p(\theta)$ is a constant, which makes ML a special case for MAP

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)$$

- Many books use MLE instead of ML for Maximum Likelihood Estimation
- ML tends to overfit.

The Log Likelihood

- Mainly for numerical reason, we work with the log posterior and likelihood functions, which are monotonically increasing.

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|\mathcal{D})$$

$$= \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D}|\theta)p(\theta)$$

$$= \underset{\theta}{\operatorname{argmax}}\, \log\big(p(\mathcal{D}|\theta)p(\theta)\big)$$

$$= \underset{\theta}{\operatorname{argmax}}[\log p(\mathcal{D}|\theta) + \log p(\theta)]$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmin}}[-\log p(\mathcal{D}|\theta) - \log p(\theta)]$$

Example: Binomial Distribution Estimation

- We flip a coin N times, and we got K heads. Assume coin flips are independent and the probability of the coin to land on a head is θ and we believe all values of θ are equally likely a priori, what would be $\hat{\theta}_{MAP}$?

N Coin flips, K Heads	
$c_i \in \{H, T\}$	Outcome of the i-th coin flip
$c_{1:N} = (c_1, c_2, \ldots, c_N)$	A sequence of all the N coin flips
$p(c = H) = \theta$ $p(c = T) = 1 - \theta$	θ is the probability of landing on a head in a flip
$\theta \sim \mathcal{U}[0,1]$	Prior of θ. It is uniformly distributed between [0,1]

Estimating Binomial Distribution Parameter

- The likelihood function is the probability of observing K heads out of N flips for a particular sequence $c_{1:N}$, given θ:

$$p(\mathcal{D}|\theta) = p(c_{1:N}|\theta) = \theta^K(1-\theta)^{N-K}$$

- The prior PDF for θ is uniform

$$p(\theta) = 1, \quad \text{for } \theta \in [0,1]$$

- Likelihood $\times$ Prior $= p(\mathcal{D}|\theta) \times p(\theta) = \theta^K(1-\theta)^{N-K}$

Learning Binomial Distribution

$$\begin{aligned}\hat{\theta}_{MAP} &= \underset{\theta}{\operatorname{argmin}}[-\log p(\mathcal{D}|\theta) - \log p(\theta)] \\ &= \underset{\theta}{\operatorname{argmin}}[-\log(\theta^K(1-\theta)^{N-K})] \\ &= \underset{\theta}{\operatorname{argmin}}[-K\log(\theta) - (N-K)\log(1-\theta)]\end{aligned}$$

$$\log(xy) = \log(x) + \log(y)$$

$$\log(x^y) = y\log(x)$$

$$\frac{d}{dx}\log(x) = \frac{1}{x}$$

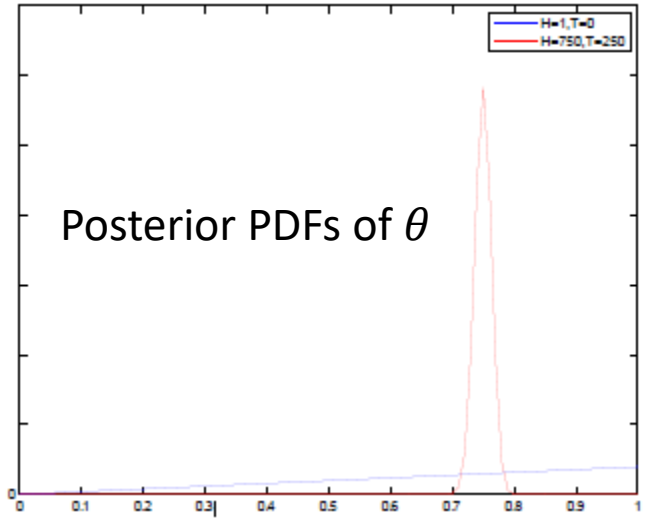
$$L(\theta) = -K\log(\theta) - (N-K)\log(1-\theta)$$

$$\frac{dL(\theta)}{d\theta} = -\frac{K}{\theta} + (N-K)\frac{1}{1-\theta} = 0$$

$$\hat{\theta}_{MAP} = \frac{K}{N}$$

Problems with MAP with Small Data Size

- A small N can lead to incorrect results
- Case: $K = 1, 2$ and we don't get any Tails.
- The posterior distribution over Heads (i.e. uncertainty of θ) is very spread out when N is small.



- There are many θ's that will give us high probability of the outcome and we only chose one of them.
- MAP is a point estimate. It does not take uncertainty into consideration.

Bayes' Estimate

- MAP finds the mode of the posterior
- Bayes' Estimate finds the mean of the posterior

$$\mathbb{E}_{p(\theta|c_{1:N})}[\theta] = \int_0^1 \theta p(\theta|c_{1:N}) d\theta = \frac{K+1}{N+2}$$

- If $N = 0$, $\hat{\theta}_{Bayes} = \frac{1}{2}$
- Compared with $\hat{\theta}_{MAP} = \frac{K}{N}$, Bayes estimate is biased from the MAP estimate towards ½.

Linear Regression Probabilistic Perspective

Linear Regression Probabilistic Model

- Multiple Regression Case: assume

$$y = \mathbf{w}^T\mathbf{x} + \eta$$

Assume Bias = 0

$$\mathbf{w} \in \mathbb{R}^D, \qquad \mathbf{x} \in \mathbb{R}^D, \qquad y \in \mathbb{R}, \qquad \eta \sim \mathcal{N}(0, \sigma^2)$$

← noise in the model

- $y \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2)$

$$p(y|\mathbf{w}, \mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y - \mathbf{w}^T\mathbf{x})^2\right)$$

Likelihood

Bayes Rule

$$p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M})p(\mathcal{M})}{p(\mathcal{D})} \qquad p(\mathbf{w}|y_{1:N}, \mathbf{x}_{1:N}) = \frac{p(y_{1:N}|\mathbf{w}, \mathbf{x}_{1:N})p(\mathbf{w}|\mathbf{x}_{1:N})}{p(y_{1:N}|\mathbf{x}_{1:N})}$$

- For the probability distribution over data, we are focusing on the output data probability density function.
- The input data is the additional information we have and is what the probability density functions are conditioned on. But it is not random.
- The weights are random variables and are what the output data probability density functions are conditioned on.

The Posterior over $\mathbf{w}$

- The Goal is to compute

$$p(\mathbf{w}|\mathbf{x}_{1:N}, y_{1:N}) = \frac{p(y_{1:N}|\mathbf{w}, \mathbf{x}_{1:N})p(\mathbf{w}|\mathbf{x}_{1:N})}{P(y_{1:N}|\mathbf{x}_{1:N})}$$

$$= \frac{p(y_{1:N}|\mathbf{w}, \mathbf{x}_{1:N})p(\mathbf{w})}{p(y_{1:N}|\mathbf{x}_{1:N})}$$

$$= \kappa p(y_{1:N}|\mathbf{w}, \mathbf{x}_{1:N})p(\mathbf{w})$$

Assume that $\mathbf{x}$ alone provides no information about $\mathbf{w}$

Evidence term does not depend on $\mathbf{w}$. It is a constant κ^{-1}.

Joint Likelihood of Output Data

- Assume noises of data points are independent from each other.
- The joint likelihood is given by

$$p(y_{1:N}|\mathbf{w}, \mathbf{x}_{1:N}) = \prod_{i=1}^{N} p(y_i|\mathbf{w}, \mathbf{x}_i)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2\right)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2\right)$$

The Prior of $\mathbf{w}$

- Assume
  - Gaussian weight decay prior for the unknown weights $\mathbf{w}$
  - Each element in $\mathbf{w}$ is IID (independently and identically distributed)
- $\mathbf{w} \sim \mathcal{N}(0, \alpha\mathbf{I})$, where $\mathbf{w} \in \mathbb{R}^D$ and $\alpha > 0$ is the variance.

$$p(\mathbf{w}) = \prod_{j=1}^{D} \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{1}{2\alpha} w_j^2\right)$$

$$= \left(\frac{1}{2\pi\alpha}\right)^{\frac{D}{2}} \exp\left(-\frac{1}{2\alpha} \sum_{j=1}^{D} w_j^2\right)$$

$$p(\mathbf{w}) = \left(\frac{1}{2\pi\alpha}\right)^{\frac{D}{2}} \exp\left(-\frac{1}{2\alpha} \mathbf{w}^T\mathbf{w}\right)$$

$$= \left(\frac{1}{2\pi\alpha}\right)^{\frac{D}{2}} \exp\left(-\frac{1}{2\alpha} \|\mathbf{w}\|^2\right)$$

$\|\mathbf{Xw} - \mathbf{y}\|^2 = \sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2$

Posterior over $\mathbf{w}$

$$p(\mathbf{w}|\mathbf{x}_{1:N}, y_{1:N}) = \kappa p(y_{1:N}|\mathbf{w}, \mathbf{x}_{1:N})p(\mathbf{w})$$

$$= \kappa \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{Xw} - \mathbf{y}\|^2\right)\left(\frac{1}{2\pi\alpha}\right)^{\frac{D}{2}} \exp\left(-\frac{1}{2\alpha}\|\mathbf{w}\|^2\right)$$

The negative log-posterior is

$$L(\mathbf{w}|\mathbf{x}_{1:N}, y_{1:N}) = -\ln p(\mathbf{w}|\mathbf{x}_{1:N}, y_{1:N})$$

constant C

$$= \left(\frac{1}{2\sigma^2}\|\mathbf{Xw} - \mathbf{y}\|^2\right) + \left(\frac{1}{2\alpha}\|\mathbf{w}\|^2\right) \underbrace{-\ln\kappa + \frac{N}{2}\ln(2\pi\sigma^2) + \frac{D}{2}(2\pi\alpha)}_{C}$$

The MAP Estimation

$$\|\mathbf{Xw}-\mathbf{y}\|^2 = \sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2$$

$$L(\mathbf{w}|\mathbf{x}_{1:N}, y_{1:N}) = \left(\frac{1}{2\sigma^2}\|\mathbf{Xw}-\mathbf{y}\|^2\right) + \left(\frac{1}{2\alpha}\|\mathbf{w}\|^2\right) + C \quad (1)$$

$$2\sigma^2 L(\mathbf{w}|\mathbf{x}_{1:N}, y_{1:N}) = \|\mathbf{Xw}-\mathbf{y}\|^2 + \frac{2\sigma^2}{2\alpha}\|\mathbf{w}\|^2 + C' \quad (2)$$

$$\begin{aligned}\underset{\mathbf{w}}{\operatorname{argmin}}\, L(\mathbf{w}|\mathbf{x}_{1:N}, y_{1:N}) &= \underset{\mathbf{w}}{\operatorname{argmin}}\, 2\sigma^2 L(\mathbf{w}|\mathbf{x}_{1:N}, y_{1:N}) \\ &= \underset{\mathbf{w}}{\operatorname{argmin}} \left(\|\mathbf{Xw}-\mathbf{y}\|^2 + \frac{2\sigma^2}{2\alpha}\|\mathbf{w}\|^2\right) \\ &= \underset{\mathbf{w}}{\operatorname{argmin}} (\|\mathbf{Xw}-\mathbf{y}\|^2 + \lambda\|\mathbf{w}\|^2)\end{aligned}$$

Remark: Ridge Regression and MAP

- Linear least squares with regularization is a form of MAP estimation.
- The hyper parameter $\lambda = \frac{\sigma^2}{\alpha}$ is the ratio of the observed variance and the variance of our belief in the prior.
- As the variance of observed data increases, we choose bigger λ
- Ridge regression assumes the Gaussian prior.
- If the prior changes, the regularization term changes.

Acknowledgement

- Prof. David Fleet developed the course. He made his notes and courseware available to all of us.
- Prof. Francisco (Paco) shared his assignments with fellow colleagues.
- Prof. Rawad A. Assi shared past assignments with me