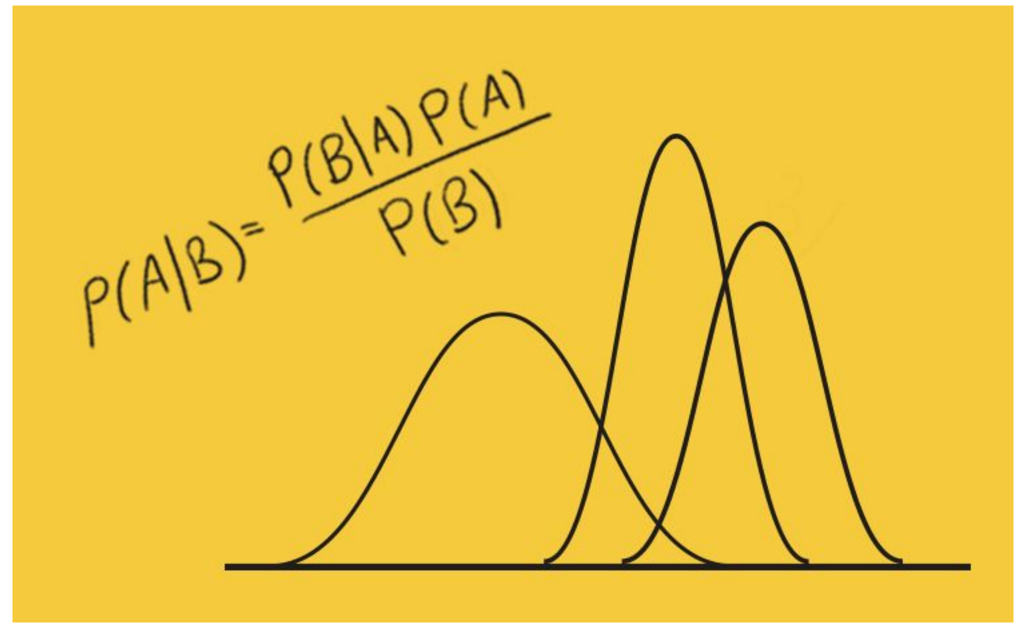


Bayesian Statistics (Basic) Cheat Sheet

V2020.10.14

(Dr Yan Xu)

Probability

- joint probability

$$\mathbb{P}(A = a_1, B = b_1)$$

- conditional probability

$$\mathbb{P}(A = a_1 | B = b_1) = \frac{\mathbb{P}(A = a_1, B = b_1)}{\mathbb{P}(B = b_1)}$$

Independence

- (absolutely) independent

$$\mathbb{P}(A = a | B = b) = \mathbb{P}(A = a)$$

- conditionally independent

$$\mathbb{P}(A = a | B = b, C = c) = \mathbb{P}(A = a | C = c)$$

- absolutely independent ***does not imply*** conditionally independent.
- also, conditionally independent ***does not imply*** absolutely independent.

Bayes Theorem

- Formula for 1 dependent variable:

$$\mathbb{P}(A = a_1 | B = b_1) = \frac{\mathbb{P}(A = a_1, B = b_1)}{\mathbb{P}(B = b_1)} = \frac{\mathbb{P}(A = a_1) \times \mathbb{P}(B = b_1 | A = a_1)}{\mathbb{P}(B = b_1)}$$

- $\mathbb{P}(A = a_1)$ is called the **prior** probability
- $\mathbb{P}(B = b_1 | A = a_1)$ is called the **likelihood**
- $\mathbb{P}(B = b_1)$ is called the **evidence**
- $\mathbb{P}(A = a_1 | B = b_1)$ is called the **posterior** probability

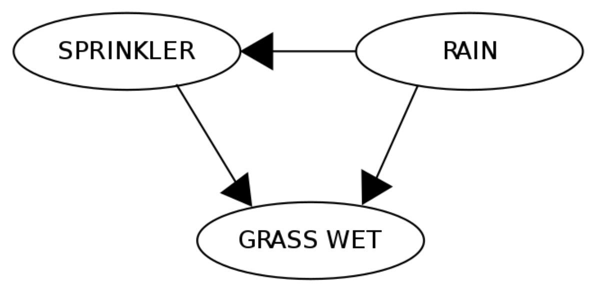
- The case for two dependent variables:

$$\mathbb{P}(A = a_1 | B = b_1, C = c_1) = \frac{\mathbb{P}(A = a_1 | C = c_1) \times \mathbb{P}(B = b_1 | A = a_1, C = c_1)}{\mathbb{P}(B = b_1 | C = c_1)}$$

Bayes Network (or Graphical Model)

- A directed acyclic graph (DAG)
  i. nodes: variables
  ii. edges: dependencies

RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99



RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

(source: wikipedia)

- determining independence: d-separation method
  i. draw the ancestral graph
  ii. connect parents of joint child
  iii. remove directions
  iv. remove given variables
  v. path -> dependence

General Posterior Distribution Inference

- Enumeration
- Elimination
- Sampling (*conditional prob -> prob*)
  - Rejection Sampling
  - Gibbs Sampling
  - Metropolis-Hasting Sampling

Conjugate Prior

- An Example:
  likelihood function: ***Binomial(N, p)***
  prior ***p***: **Beta(*α*, β)**
  posterior ***p|X*** : **Beta(*α*+x, β+(n-x))**
- If likelihood function is discrete:

Likelihood distribution	Conjugate prior
Binomial	Beta
Poisson	Gamma
Geometric	Gamma
Multinomial	Dirichlet

- If likelihood function is continuous:

Likelihood distribution	Conjugate prior
Uniform	Pareto
Exponential	Gamma
Normal distribution, known σ^2	Normal distribution
Normal distribution, known μ	Inverse Gamma

Common Bayesian ML Models

- Naive Bayes
- Bayesian Linear Regression
- Bayesian Logistic Regression
- Linear/Quadratic Discriminant Analysis
- Gaussian Mixture Model (or Bayesian K-mean)
- Latent Dirichlet Allocation (topic model)

Naive Bayes (Classifiers)

- Definition:

Suppose Y represents the class variable and $X_1, X_2, X_3, \ldots X_n$ are inputs:

$$P(Y|X_1, \ldots, X_n) = \frac{P(X_1, \ldots, X_n|Y)\ P(Y)}{P(X_1, \ldots, X_n)}$$

Assuming **$X_i \perp X_j$ given Y** for all i, j , we may write the above equation as:

$$P(Y|X_1, \ldots, X_n) = \frac{P(X_1|Y)\, P(X_2|Y) \ldots P(X_n|Y)\ P(Y)}{P(X_1, \ldots, X_n)}$$ **(Naïve Assumption)**

- Why naive?
  assuming features are *conditionally independent*
- Prediction:

$$\hat{Y} = argmax_y P(Y) \prod_{i=1}^{n} P(X_i|Y)$$

- Variants (depending on P(X|Y))
  - Multinomial Naive Bayes
    - prior/posterior: Dirichlet
  - Gaussian Naive Bayes
    - prior/posterior: Gaussian
- Parameters could be estimated using MLE (e.g. sklearn) or MAP.
- For sklearn, the probability outputs are not to be taken too seriously
- Cannot naturally deal with mixed categorical and continuous features

Bayesian Linear Regression

- Likelihood Function:

$$J(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2$$

- Prior:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{V}_0)$$

- Posterior:

$$\begin{aligned} p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) &\propto \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{V}_0)\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N) = \mathcal{N}(\mathbf{w}|\mathbf{w}_N, \mathbf{V}_N) \\ \mathbf{w}_N &= \mathbf{V}_N \mathbf{V}_0^{-1} \mathbf{w}_0 + \frac{1}{\sigma^2} \mathbf{V}_N \mathbf{X}^T \mathbf{y} \\ \mathbf{V}_N &= \sigma^2 (\sigma^2 \mathbf{V}_0^{-1} + \mathbf{X}^T \mathbf{X})^{-1} \end{aligned}$$

Bayesian Linear Regression (continued)

- L2 Regularization
  If $\mathbf{w}_0 = \mathbf{0}$ and $\mathbf{V}_0 = \boldsymbol{\tau}^2 \boldsymbol{I}$, then the posterior mean reduces to ***L2*** estimation, with $\lambda = \boldsymbol{\sigma}^2/\boldsymbol{\tau}^2$ *(or noise / prior)*
- Prediction: (mean & variance)

$$\begin{aligned} p(y|\mathbf{x}, \mathcal{D}, \sigma^2) &= \int \mathcal{N}(y|\mathbf{x}^T\mathbf{w}, \sigma^2)\mathcal{N}(\mathbf{w}|\mathbf{w}_N, \mathbf{V}_N) d\mathbf{w} \\ &= \mathcal{N}(y|\mathbf{w}_N^T \mathbf{x}, \sigma_N^2(\mathbf{x})) \\ \sigma_N^2(\mathbf{x}) &= \sigma^2 + \mathbf{x}^T \mathbf{V}_N \mathbf{x} \end{aligned}$$

- The observation noise may dominate variance in case of large data

Bayesian Logistic Regression

- Negative Log-likelihood (Cross-Entropy):

$$f(w) = \sum_{i=1}^{N} \{y_i * log(p_i) + (1 - y_i) * log(1 - p_i)\}$$

$$p_i = \frac{e^{Xw}}{1 + e^{Xw}}$$

- Unlike Linear Regression, there is no conjugate prior for Logistic Regression
- Prior: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{V}_0)$
- Laplace/Gaussian Approximation:
  posterior approximated:

$$f(w) = \sum_{i=1}^{N} \{y_i * log(p_i) + (1 - y_i) * log(1 - p_i)\} + \frac{1}{2}(w - w_0)^T V_0^{-1} (w - w_0)$$

- Posterior (mean)

$$\hat{w} = argmin_w f(w)$$

- Luckily, f(x) is convex!! So, L-BFGS.

$$\triangledown f(w) = X^T(P - Y) + V^{-1}(W - W_0)$$

- Posterior (Variance)

$$V_N^{-1} = \triangledown^2 f(w) = V_0^{-1} + \sum_{i=1}^{N} p_i * (1 - p_i) X X^T$$

you may see *Hessian* (inverse covariance) instead of *V*.

- Prediction: (mean)
- Prediction: (variance) - Sampling
- ML Example