CS 559 Machine Learning

Introduction and Overview

Yue Ning
Department of Computer Science
Stevens Institute of Technology

Probability

Discrete Random Variable

- Sample space Ω: Possible "states" x of the random variable X (outcomes of the experiment, output of the system, measurement).
- Discrete random variables either have a finite or countable number of states.
- Events: Possible combinations of states ('subsets of Ω')

Discrete Random Variable

- Probability mass function $P(X = x)$: a function which tells us how likely each possible outcome is.

$$P(X = x) = P_X(x) = P(x) \tag{1}$$

$$P(x) \geq 0 \text{ for each } x \tag{2}$$

$$\sum_{x \in \Omega} P(x) = 1 \tag{3}$$

$$P(A) = P(x \in A) = \sum_{x \in A} P(X = x) \tag{4}$$

- We write: $X|q \sim \text{Binomial}(q)$
- Bernoulli, Binomial, Multinonomial, Poisson

Conditional Probability

- Conditional probability: Recalculated probability of event A after someone tells you that event B happened.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5}$$

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \tag{6}$$

- Bayes Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{7}$$

Expectation and Variance

- Expectation (or mean): $E(X) = \sum_x P(X = x)x$
- Expectation of a function: $E(f(X)) = \sum_x P(X = x)f(x)$
- Moments = expectation of power of X: $M_k = E(X^k)$
- Variance: Average (squared) fluctuation from the mean

$$\begin{aligned} \mathrm{Var}(X) &= E((X - E(X))^2) && (8) \\ &= E(X^2) - E(X)^2 && (9) \\ &= M_2 - M_1^2 && (10) \end{aligned}$$

- Standard deviation: Square root of variance. Aside: Difference between expectation/variance of random variable and empirical average/variance.

Bivariate Distributions

- Joint distribution: $P(X = x, Y = y)$, a list of all probabilities of all possible pairs of observations
- Marginal distribution: $P(X = x) = \sum_y P(X = x, Y = y)$
- Conditional distribution: $P(X = x|Y = y) = \frac{P(X=x, Y=y)}{P(y=y)}$
- $X|Y$ has distribution $P(X|Y)$, where $P(X|Y)$ specifies a "lookup-table" of all possible $P(X = x|Y = y)$

Conditioning and marginalization come up in Bayesian inference ALL the time: Condition on what you observe. Marginalize out the uncertainty.

Expectation and Covariance of Multivariate Distributions

- Conditional distributions are just distributions which have a (conditional) mean or variance.
- Covariance is the expected value of the product of fluctuations:

$$
\begin{aligned}
\mathrm{Cov}(X, Y) &= E\left((X - E(X))(Y - E(Y))\right) && (11)\\
&= E(XY) - E(X)E(Y) && (12)\\
\mathrm{Var}(\mathrm{X}) &= \mathrm{Cov}(X, X) && (13)
\end{aligned}
$$

Aside: One common way to construct bivariate random variables is to have a random variable whose parameter is another random variable.

Independence of Random Variables

- Intuitively, two events are independent if knowing that the first took places tells us nothing about the probability of the second: $P(A|B) = P(A)$
- $P(A)P(B) = P(A \cap B)$
- Two random variables are independent if the joint p.m.f. is the product of the marginals: $P(X = x, Y = y) = P(X = x)P(Y = y)$.
- If X and Y are independent, we write $X \perp Y$. Knowing the value of X does not tell us anything about Y.
- If X and Y are independent, $\mathrm{Cov}(X, Y) = 0$.

Aside: Mutual information is a measure of how "non-independent" two random variables are.

Multivariate Distributions

- $\mathbf{X}, \mathbf{x}$ are vector valued.
- Mean: $E(\mathbf{X}) = \sum_{\mathbf{x}} \mathbf{x} P(\mathbf{x})$
- Covariance matrix:

$$\mathrm{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) \tag{14}$$

$$\mathrm{Cov}(\mathbf{X}) = E(\mathbf{X}\mathbf{X}^\top) - E(\mathbf{X})E(\mathbf{X})^\top \tag{15}$$

- Conditional and marginal distributions: Can define and calculate any (multi or single-dimensional) marginals or conditional distributions we need: $P(X_1)$, $P(X_1, X_2)$, $P(X_1, X_2, X_3 | X_4)$, etc..

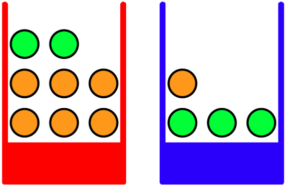
Example (from Bishop. [2006])

Assuming, we know that:

$$P(B = r) = 4/10 \tag{16}$$
$$P(B = b) = 6/10 \tag{17}$$

The probability of selecting a fruit from a given box is:

$$P(F = a|B = r) = 1/4 \tag{18}$$
$$P(F = o|B = r) = 3/4 \tag{19}$$
$$P(F = a|B = b) = 3/4 \tag{20}$$
$$P(F = o|B = b) = 1/4 \tag{21}$$

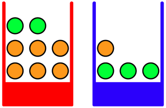


Example (from Bishop. [2006])

- What is the probability of choosing an apple?
- Conditional probability:

$$
\begin{aligned}
P(F=a) &= P(F=a|B=r)P(B=r) \\
&\quad + P(F=a|B=b)P(B=b) \qquad (22) \\
&= \frac{1}{4}\times\frac{4}{10}+\frac{3}{4}\times\frac{6}{10} = 11/20 \qquad (23)
\end{aligned}
$$

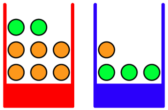
- Thus, the probability of choosing an orange is $P(F=o) = 1 - 11/20 = 9/20$



Example (from Bishop. [2006])

We are told that a piece of fruit has been selected and it is an orange, and we would like to know which box it came from. Using Bayes' theorem,

$$P(B = r|F = o) = \frac{P(F = o|B = r)P(B = r)}{P(F = o)} = \frac{\frac{3}{4} \times \frac{4}{10}}{\frac{9}{20}} = 2/3 \tag{24}$$



Prior vs. Posterior

STEVENS

Prior Probability

If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability $P(B)$.

Posterior Probability

Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $P(B|F)$, which we shall call the posterior probability because it is the probability obtained after we have observed F.

Bayesian Probabilities

Bayesian view: probabilities provide a quantification of uncertainty. Before observing the data, the assumptions about w are captured in the form of a prior probability distribution $P(\mathbf{w})$. The effect of the observed data $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ is expressed by $P(\mathcal{D}|\mathbf{w})$.

Bayes' theorem:

$$P(\mathbf{w}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathbf{w})P(\mathbf{w})}{P(\mathcal{D})}$$

Bayes' theorem in words:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Continuous random variables

- A random variable X is continuous if its sample space X is uncountable.
- In this case, $P(X = x) = 0$ for each x.
- If $p_X(x)$ is a probability density function for X, then

$$P(a < X < b) = \int_a^b p(x)dx \tag{25}$$

$$P(a < X < a + dx) \approx p(a) \cdot dx \tag{26}$$

- The cumulative distribution function is $F_X(x) = P(X < x)$. We have that $p_X(x) = F'(x)$, and $F(x) = \int_{-\infty}^{x} p(s)ds$.

Continuous random variables (cont.)

- More generally: If A is an event, then

$$P(A) = P(X \in A) = \int_{x \in A} p(x)dx \tag{27}$$

$$P(\Omega) = P(X \in \Omega) = \int_{x \in \Omega} p(x)dx = 1 \tag{28}$$

- Example: Uniform, Exponential, Beta [on board]

Two sloppy facts: Probability vs. Probability Density

- $P(X = x)$: "the probability of X" when they really mean "the probability density of X evaluated at x".
- "We need to integrate out X": when X are discrete random variables, the integrals would need to be replaced by sums.
- It is usually clear from the context whether a random variable is discrete or continuous.

Mean, Variance and Conditionals

- Mean: $E(X) = \int_x x \cdot p(x)dx$
- Variance: $\text{Var}(X) = E(X^2) - E(X)^2$
- Example: Uniform [quiz]
- If X has pdf $p(x)$, then $X|(X \in A)$ has pdf

$$p_{X|A}(x) = \frac{p(x)}{P(A)} = \frac{p(x)}{\int_{x\in A} p(x)dx} \tag{29}$$

- Only makes sense if $P(A) > 0$!

Bivariate Continuous Distributions

- $p_{X,Y}(x, y)$, joint probability density function of X and Y
- $\int_x \int_y p(x, y) dx dy = 1$
- Marginal distribution: $p(x) = \int_{-\infty}^{\infty} p(x, y) dy$
- Conditional distribution: $p(x|y) = \frac{p(x,y)}{p(y)}$
- Note: $P(Y = y) = 0$!
- Independence: X and Y are independent if $p_{X,Y}(x, y) = p_X(x) p_Y(y)$

The Univariate Gaussian

- Probability density function

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{30}$$

- Easy to validate:

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)dx = 1 \tag{31}$$

- Expectation

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)xdx = \mu \tag{32}$$

- Variance

$$var[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \tag{33}$$

Products of Gaussian pdfs

- Suppose $p_1(x) = \mathcal{N}(x, \mu_1, 1/\beta_1)$ and $p_2(x) = \mathcal{N}(x, \mu_2, 1/\beta_2)$

$$p_1(x)p_2(x) \propto \mathcal{N}(x, \mu, 1/\beta) \tag{34}$$

$$\beta = \beta_1 + \beta_2 \tag{35}$$

$$\mu = \frac{1}{\beta}(\beta_1\mu_1 + \beta_2\mu_2) \tag{36}$$

In general:

$$p_1(x)p_2(x)...p_n(x) \propto \mathcal{N}(x, \mu, 1/\beta) \tag{37}$$

$$\beta = \sum_n \beta_n \tag{38}$$

$$\mu = \frac{1}{\beta}\sum_n \mu_n\beta_n \tag{39}$$

This is also true for multivariate Gaussians!

Maximum Likelihood (ML) estimator

- Assuming data points are independent and identically distributed (i.i.d.), the probability of the data set given μ and σ^2 (the likelihood function):

$$P(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) \tag{40}$$

- Log-likelihood:
- Maximizing Log-likelihood with respect to μ and σ^2:

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{41}$$

$$\sigma^2_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 \tag{42}$$

Maximum Likelihood (ML) estimator

- The ML solutions μ_{ML} and σ^2_{ML} are functions of the data set values $x_1, .., x_N$. The expectations of these quantities w.r.t the data set values:

$$\mathbb{E}[\mu_{ML}] = \mu \tag{43}$$

$$\mathbb{E}[\sigma^2_{ML}] = \left(\frac{N-1}{N}\right)\sigma^2 \tag{44}$$

- The ML estimator obtains correct means but underestimate the true variance by a factor $\frac{N-1}{N}$

Maximum a posteriori probability (MAP) estimator

- Given input values $\mathbf{x} = (x_1, ..., x_N)^T$ and their corresponding target values $\mathbf{y} = (y_1, ..., y_N)^T$.
- Express our uncertainty over the values of the target variables:

$$p(y|x, \mathbf{w}, \beta) = \mathcal{N}(y|f(x, \mathbf{w}), \beta^{-1})$$

- Using our training data to determine the unknown parameters $\mathbf{w}, \beta$ by maximum likelihood:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|f(x_n, \mathbf{w}), \beta^{-1})$$

- Log Likelihood:

$$lnp(y|x, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}(f(x_n, \mathbf{w}) - y_n)^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

MAP estimator

- A more Bayesian approach by introducing a prior distribution for **w**:

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-\mathbf{1}}\mathbf{I}) = (\frac{\alpha}{\mathbf{2}\pi})^{\mathbf{(M+1)/2}} \exp(-\frac{\alpha}{\mathbf{2}}\mathbf{w}^{\mathbf{T}}\mathbf{w})$$

- Using Bayes' theorm, the posterior distribution for w:

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta) \propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

- Taking the negative logarithm, Maximum posterior (MAP) is equivalent to minimizing the regularized sum-of-squares error function:

$$\frac{\beta}{2}\sum_{n=1}^{N}\{f(x_n, \mathbf{w}) - y_n)\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \tag{45}$$

Bayesian Probabilities

- A key issue in pattern recognition is uncertainty. It is due to incomplete and/or ambiguous information, i.e. finite and noisy data.
- Probability theory and decision theory provide the tools to make optimal predictions given the limited available information.
- In particular, the Bayesian interpretation of probability allows to quantify uncertainty, and make precise revisions of uncertainty in light of new evidence.

Readings

- Bishop Chapter 1&2
- For next week: Bishop Chapter 3