CS542
Neural Computation With
Artificial Neural Networks

Lecture 02
Statistics Refresher
Outline

- Random Variables
  - Discrete & continuous
- Distributions
  - Discrete & continuous
- Expected values and moments
- Joint distributions, conditional distributions, independence
- Reading assignment for next class
  - Bishop Chapter 1
Random Variables

- A random variable is a random number determined by chance, or, more formally, drawn according to a probability distribution
  - The probability distribution can be given by the physics of an experiment (e.g., throwing dice)
  - The probability distribution can be synthetic
  - Discrete & continuous random variables exist

- Typical random variables in statistical learning
  - Input data, output data, noise

- Important concept: The data generating model
  - E.g., what is the data generating model for i) throwing dice, ii) regression, iii) classification, iv) visual perception?

- Problem: On which time scale is a distribution observed?
Discrete Distribution

- The random variables only take on discrete values
  - E.g, throwing dice: possible values are: \( v_i \in \{1,2,3,4,5,6\} \)

- The probabilities sum to 1:

\[
\sum P(v_i) = 1
\]

- Discrete distributions are particularly important in classification and decision making

- Probability mass function or frequency function is a normalized histogram
Bernoulli Distribution

- A Bernoulli random variable only takes on two values, e.g., 0 and 1
  \[ p(x = 1 | \mu) = \mu \]
  \[ \text{Bern}(x | \mu) = \mu^x (1 - \mu)^{1-x} \]
  \[ \mathbb{E}[x] = \mu \]
  \[ \text{var}[x] = \mu (1 - \mu) \]

- Bernoulli distributions are naturally modeled with sigmoidal nonlinearities in statistical learning
Binomial Distribution

- Binomial variables are a sequence of $N$ repeated Bernoulli variables

$$p(m \text{ heads}|N, \mu)$$

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m \text{Bin}(m|N, \mu) = N\mu$$

$$\text{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \text{Bin}(m|N, \mu) = N\mu(1 - \mu)$$

- Binomial variables are important for, e.g., density estimation: “What is the probability that $k$ out of $n$ data points fall into region $R$?”

- The Bernoulli distribution is a subset of the Binomial distribution ($N=1$)
Binomial Distribution

Bin(m|10, 0.25)
Multinomial Variables

- Multinomial variables are a generalization of binomial variables to multiple outputs (e.g., multiple classes)

1-of-K coding scheme: \( \mathbf{x} = (0, 0, 1, 0, 0, 0)^T \)

\[
p(\mathbf{x} | \mu) = \prod_{k=1}^{K} \mu_k^{x_k}
\]

\( \forall k : \mu_k \geq 0 \) and \( \sum_{k=1}^{K} \mu_k = 1 \)

\[
\mathbb{E}[\mathbf{x} | \mu] = \sum_{\mathbf{x}} p(\mathbf{x} | \mu) \mathbf{x} = (\mu_1, \ldots, \mu_K)^T = \mu
\]

\[
\sum_{\mathbf{x}} p(\mathbf{x} | \mu) = \sum_{k=1}^{K} \mu_k = 1
\]
The Multinomial Distribution

- N independent trials can result in one of K types of outcome
- What is the probability that in N trials, the frequency of the K classes is $m_1, m_2, \ldots, m_k$

$$\text{Mult}(m_1, m_2, \ldots, m_K | \mu, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

$$\mathbb{E}[m_k] = N \mu_k$$

$$\text{var}[m_k] = N \mu_k (1 - \mu_k)$$

$$\text{cov}[m_j m_k] = -N \mu_j \mu_k$$
The Multinomial Distribution

- The multinomial distribution plays an important role in multi-class classification (N=1)

To which class does a data vector belong?
The Poisson distribution is a binomial distribution where the number of trials $N$ goes to infinity, and the probability of success on each trial, $\pi$, goes to zero, such that $N \pi = \lambda$.

$P(m) = \frac{\lambda^m}{m!} e^{-\lambda}$

For example, Poisson distributions are an important model for the firing characteristics of biological neurons. They are also used as an approximation to binomial variables with small $p$. 
Poisson Distribution

- Example: What is the probability of firing of a Purkinje neuron in the cerebellum in a 10ms time interval?
  - We now that the average firing of these neurons is about 40Hz
  - $\lambda = 40\text{Hz} \times 0.01\text{s}$
  - Note that this approximation only work if the number of spike is low in the given time interval
Continuous Distribution

- Random variables take on real values
- Continuous distributions are discrete distributions where the number of discrete values goes to infinity, while the probability of each value goes to zero.
- Thus probabilities become densities.
- Probability densities integrate to 1
  \[ \int_{-\infty}^{+\infty} p(x) \, dx = 1 \]
- Continuous distributions are particularly important in regression and unsupervised learning
Continuous Distribution

- Example of a density $p(x)$

$$P(a < x < b) = \int_a^b p(x) \, dx$$
The Gaussian Distribution

\[ N(x | \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} \]

\[ N(x | \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \]
Central Limit Theorem

- The distribution of the sum of $N$ i.i.d. random variables becomes increasingly Gaussian as $N$ grows.
- Example: $N$ uniform $[0,1]$ random variables.
Geometry of the Multivariate Gaussian

\[ \Delta^2 = (x - \mu)^T \Sigma^{-1} (x - \mu) \]

Mahalanobis distance
(very important in multi-variate distances)

\[ \Sigma^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} u_i u_i^T \]

\[ \Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \]

\[ y_i = u_i^T (x - \mu) \]
Moments of the Multivariate Gaussian

\[
\mathbb{E}[x] = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \int \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} x \, dx \\
= \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \int \exp \left\{ -\frac{1}{2} z^T \Sigma^{-1} z \right\} (z + \mu) \, dz
\]

thanks to anti-symmetry of z

\[
\mathbb{E}[x] = \mu
\]
Moments of the Multivariate Gaussian

\[ \mathbb{E}[xx^T] = \mu \mu^T + \Sigma \]

\[ \text{cov}[x] = \mathbb{E} \left[ (x - \mathbb{E}[x])(x - \mathbb{E}[x])^T \right] = \Sigma \]
Partitioned Gaussian Distributions

\[ p(x) = \mathcal{N}(x | \mu, \Sigma) \]

\[ x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \]

\[ \Lambda \equiv \Sigma^{-1} \quad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \]
Partitioned Conditionals and Marginals

\[ p(x_a | x_b) = \mathcal{N}(x_a | \mu_{a|b}, \Sigma_{a|b}) \]

\[ \Sigma_{a|b} = \Lambda_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba} \]

\[ \mu_{a|b} = \Sigma_{a|b} \{ \Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b) \} \]

\[ = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab} (x_b - \mu_b) \]

\[ = \mu_a + \Sigma_{ab} \Sigma_{bb}^{-1} (x_b - \mu_b) \]

\[ p(x_a) = \int p(x_a, x_b) \, dx_b \]

\[ = \mathcal{N}(x_a | \mu_a, \Sigma_{aa}) \]
Partitioned Conditionals and Marginals

\[ x_b = 0.7 \]

\[ p(x_a, x_b) \]

\[ p(x_a | x_b = 0.7) \]

\[ p(x_a) \]
The Exponential Family

- A large class of distributions that are all analytically appealing. Why? Because taking the log() of them decomposes them into simple terms.

- All distributions are uni-modal

\[
p(x|\eta) = h(x) g(\eta) \exp \left\{ \eta^T u(x) \right\}
\]

- where \( \eta \) is the natural parameter and

\[
g(\eta) \int h(x) \exp \left\{ \eta^T u(x) \right\} \, dx = 1
\]

- so \( g(\eta) \) can be interpreted as a normalization coefficient.
The Exponential Family
Bernoulli Distribution

- The Bernoulli Distribution

\[
p(x|\mu) = \text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x} \\
= \exp \left\{ x \ln \mu + (1 - x) \ln (1 - \mu) \right\} \\
= (1 - \mu) \exp \left\{ \ln \left( \frac{\mu}{1 - \mu} \right) x \right\}
\]

- Comparing with the general form we see that

\[
\eta = \ln \left( \frac{\mu}{1 - \mu} \right) \quad \text{and so} \quad \mu = \sigma(\eta) = \frac{1}{1 + \exp(-\eta)}.
\]

Logistic sigmoid
The Exponential Family
Bernoulli Distribution

- The Bernoulli distribution can hence be written as

\[ p(x | \eta) = \sigma(-\eta) \exp(\eta x) \]

- where

\[
\begin{align*}
u(x) & = x \\
h(x) & = 1 \\
g(\eta) & = 1 - \sigma(\eta) = \sigma(-\eta).
\end{align*}
\]
The Exponential Family
Multinomial Distribution

- The Multinomial Distribution

\[
p(x|\mu) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp \left\{ \sum_{k=1}^{M} x_k \ln \mu_k \right\} = h(x) g(\eta) \exp (\eta^T u(x))
\]

- where

\[
x = (x_1, \ldots, x_M)^T \quad \eta = (\eta_1, \ldots, \eta_M)^T
\]

\[
\eta_k = \ln \mu_k \\
u(x) = x \\
h(x) = 1 \\
g(\eta) = 1.
\]

NOTE: The \(\mu_k\) parameters are not independent since the corresponding \(x_k\) must satisfy

\[
\sum_{k=1}^{M} \mu_k = 1.
\]
The Exponential Family
Multinomial Distribution

• Let $\mu_M = 1 - \sum_{k=1}^{M-1} \mu_k$. This leads to

• and

$\eta_k = \ln \left( \frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j} \right)$

$\mu_k = \frac{\exp(\eta_k)}{1 + \sum_{j=1}^{M-1} \exp(\eta_j)}$.

• Here the $\mathbb{W}_k$ parameters are independent. Note that

• and

$0 \leq \mu_k \leq 1$

$\sum_{k=1}^{M-1} \mu_k \leq 1$. 
The Exponential Family
Multinomial Distribution

- The Multinomial distribution can then be written as

\[ p(x|\mu) = h(x)g(\eta) \exp(\eta^T u(x)) \]

- where

\[ \eta = (\eta_1, \ldots, \eta_{M-1}, 0)^T \]
\[ u(x) = x \]
\[ h(x) = 1 \]
\[ g(\eta) = \left( 1 + \sum_{k=1}^{M-1} \exp(\eta_k) \right)^{-1} \]
The Exponential Family
Gaussian Distribution

- The Gaussian Distribution

\[ p(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} \]

\[ = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} \mu^2 \right\} \]

- where

\[ \eta = \begin{pmatrix} \mu / \sigma^2 \\ -1 / 2\sigma^2 \end{pmatrix} \quad h(x) = (2\pi)^{-1/2} \]

\[ u(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \quad g(\eta) = (-2\eta_2)^{1/2} \exp \left( \frac{\eta_1^2}{4\eta_2} \right). \]
Uniform Distribution

- All data is equally probable within a bounded region $R$.

$$p(x) = \frac{1}{R}$$

- Uniform distribution play an important role in entropy methods and information theory.
Expectations

\[ \mathbb{E}[f] = \sum_x p(x) f(x) \]

\[ \mathbb{E}[f] = \int p(x) f(x) \, dx \]

\[ \mathbb{E}_x[f|y] = \sum_x p(x|y) f(x) \]

Conditional Expectation (discrete)

\[ \mathbb{E}[f] \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n) \]

Approximate Expectation (discrete and continuous) (what happened to probabilities?)
Expectations

Example: What is the expectation of the following distribution?
Expectations of Functions

- \( E\{g(x)\} = ? \)
  - General approach: Solve \( \int g(x) p(x) \, dx \)

- Note: in general:
  \[ E\{g(x)\} \neq g(E\{x\}) \]

- Other rules:
  \[
  \begin{align*}
  E\{ax\} &= a E\{x\} \\
  E\{x + y\} &= E\{x\} + E\{y\} \\
  E\left\{ \sum_i a_i x_i \right\} &= \sum_i a_i E\{x_i\} \\
  E\{xy\} &= E\{x\} E\{y\}
  \end{align*}
  \]
Variances and Covariances

- Variances give a measure of dispersion, while covariances give a measure of correlation

\[ \text{var}[f] = \mathbb{E} \left[ (f(x) - \mathbb{E}[f(x)])^2 \right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \]

\[ \text{cov}[x, y] = \mathbb{E}_{x,y} \left[ \{x - \mathbb{E}[x]\} \{y - \mathbb{E}[y]\} \right] \]
\[ = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x] \mathbb{E}[y] \]

\[ \text{cov}[x, y] = \mathbb{E}_{x,y} \left[ \{x - \mathbb{E}[x]\}\{y^T - \mathbb{E}[y^T]\} \right] \]
\[ = \mathbb{E}_{x,y}[xy^T] - \mathbb{E}[x] \mathbb{E}[y^T] \]

Note the very important rule:

\[ \mathbb{E}[xx^T] = \mu \mu^T + \Sigma \]
Moments of Random Variables

- Definition of a Moment
  \[ m_n = E\left\{x^n\right\} \]

- Definition of a Central Moment
  \[ cm_n = E\left\{(x - \mu)^n\right\} \]

- Useful Moments:
  - \( m_1 = \text{Mean} \)
  - \( cm_2 = \text{Variance} \)
  - \( cm_3 = \text{Skewness (measure of asymmetry)} \)
  - \( cm_4 = \text{Kurtosis (measure of heavy tailed-ness and light tailed-ness)} \)
Basic Rules of Probability Theory

- Joint Distribution $p(x, y)$
- Marginal Distribution $p(y) = \int p(x, y) \, dx$
- Conditional Distribution $p(y \mid x) = \frac{p(x, y)}{p(x)}$
- Probabilistic Independence $p(x, y) = p(x)p(y)$
- Chain Rule of Probabilities

$$p(x_1, x_2, \ldots, x_n) = p(x_1 \mid x_2, \ldots, x_n) \cdot p(x_2 \mid x_3, \ldots, x_n) \cdots p(x_{n-1} \mid x_n)p(x_n)$$
Bayes Rule

\[ p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \]

\[ p(X) = \sum_Y p(X|Y)p(Y) \]

posterior \(\propto\) likelihood × prior