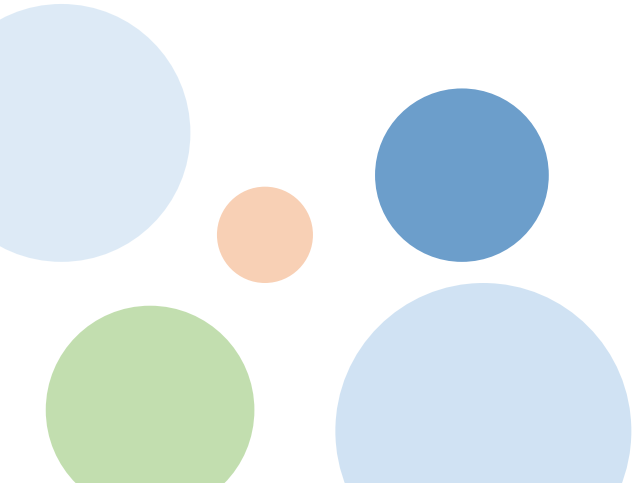
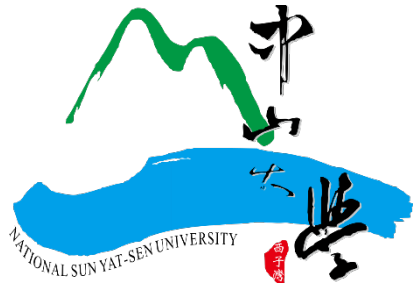
Probability





Outlines (probability)

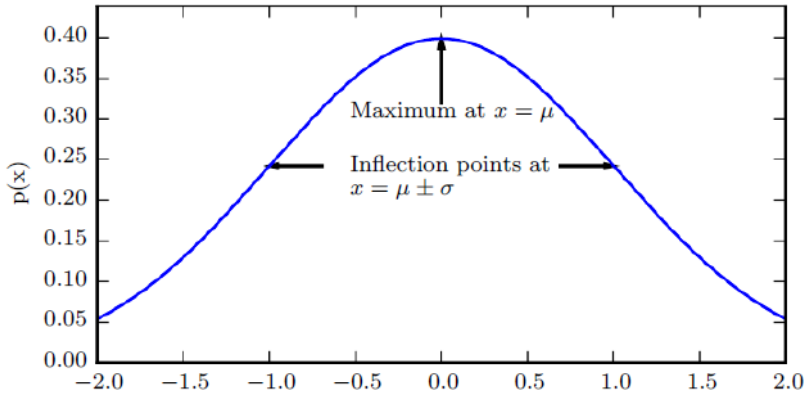
- probability distribution
- marginal probability
- conditional probability
- independence
- expectation, variance, and covariance
- Popular probability distributions
- Bayer’s rule
- information theory
- maximum likelihood estimation (MLE)

probability distribution

- random variable (r.v.)
  - a variable that can take on different values randomly
  - could be discrete or continuous, depending on the value of the outcome
- probability distribution
  - Probability Mass Function (PMF) *P(x)* for discrete r.v.
    - $0 \le P(x) \le 1, \quad \sum_{x \in \mathrm{X}} P(x) = 1$
    - e.g. uniform distribution for a discrete r.v. with *k* difference states

      $P(\mathrm{x} = x_i) = 1/k, \quad \sum_i P(\mathrm{x} = x_i) = \sum_i 1/k = k \times (1/k) = 1$
  - Probability Density Function (PDF) p(x) for continuous r.v.
    - $p(x) \ge 0, \quad \int p(x)dx = 1$
    - e.g., normal distribution (Gaussian distribution)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Common distributions

- Bernoulli(ϕ): $p(x) = \begin{cases} \phi & \text{if } x = 1 \\ 1-\phi & \text{if } x = 0 \end{cases}, \quad 0 \le \phi \le 1$
  - one toss of a coin
- binomial (n, ϕ) : $p(x) = \binom{n}{x} \phi^x (1-\varphi)^{n-x}$
  - n independent tosses of coins with x times “head”
- uniform (k): $p(x) = \begin{cases} 1/k & \text{if } x = 1, 2, \cdots, k \\ 0 & \text{otherwise} \end{cases}$

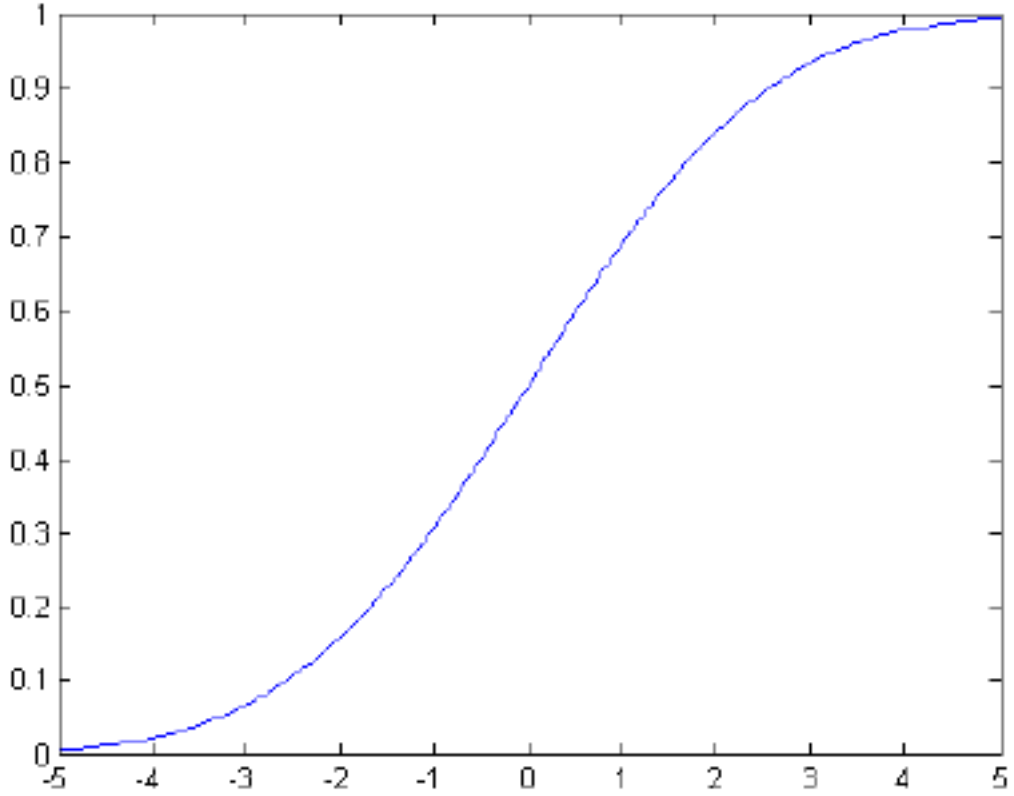
- uniform(a, b): $p(x) = \begin{cases} 1/(b-a) & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$

- Normal(μ, σ^2) $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$

cumulative distribution function (CDF)

- a function F_X : R -> [0; 1] which specifies a probability measure as,

$$F_x(x) = P(x \leq x)$$



marginal and conditional probability

- **joint** probability distribution
  - probability distribution over multiple r.v.
    - discrete r.v.: $P(\mathrm{x} = x, \mathrm{y} = y, \mathrm{z} = z)$
    - continuous r.v.: $p(x, y, z)$
- **marginal** probability distribution
  - probability distribution over a subset of random variables
    - discrete r.v. $P(\mathrm{x} = x) = \sum_y P(\mathrm{x} = x, \mathrm{y} = y)$
    - continuous r.v. $p(x) = \int p(x, y)dy$
- **conditional** probability
  - probability of some event, given that some other events has happened
  - $$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{x} = x, \mathrm{y} = y)}{P(\mathrm{x} = x)}$$

    $$\Rightarrow P(\mathrm{x} = x) \times P(\mathrm{y} = y \mid \mathrm{x} = x) = P(\mathrm{x} = x, \mathrm{y} = y)$$
  - defined only for $P(\mathrm{x} = x) > 0$

Example: conditional probability

Joint Distribution

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W = s|T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{0.2}{0.5} = 0.4$$

$$= P(W = s, T = c) + P(W = r, T = c)$$
$$= 0.2 + 0.3 = 0.5$$

Conditional Distributions

$$P(W|T)$$

$$P(W|T = hot)$$

W	P
sun	0.8
rain	0.2

$$P(W|T = cold)$$

W	P
sun	0.4
rain	0.6

Chain rule of conditional probability

- Any joint probability distribution over many random variables may be decomposed into conditional distributions over only one variable

$$P(x_1, \cdots, x_n) = P(x_1) \times \prod_{i=2}^{n} P(x_i \mid x_1, \cdots, x_{i-1})$$

- e.g.,

$$P(a,b,c) = P(a|b,c) \times P(b,c)$$

$$P(b,c) = P(b|c) \times P(c)$$

$$P(a,b,c) = P(a \mid b,c) \times P(b \mid c) \times P(c)$$

Independence

- two r.v. *x* and *y* are independent , if

$$x \perp y : P(x, y) = P(x)P(y), \quad \text{or} \quad P(x \mid y) = P(x)$$

- Two random variables x and y are **independent** if their probability distribution can be expressed as a product of two factors, one involving only x and one involving only y

$$\mathrm{x} \perp \mathrm{y} \qquad p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y).$$

- Two random variables x and y are **conditionally independent** given a random variable z if the conditional probability distribution over x and y factorizes in this way for every value of z

$$\mathrm{x} \perp \mathrm{y} \mid \mathrm{z}$$

$$p(\mathrm{x} = x, \mathrm{y} = y \mid \mathrm{z} = z) = p(\mathrm{x} = x \mid \mathrm{z} = z)p(\mathrm{y} = y \mid \mathrm{z} = z)$$

expectation

- The **expectation** or expected value of some function *f(x)* with respect to a probability distribution *P(x)* is the average or mean value that f takes on when x is drawn from P:

$$E_{x\sim P}[f(x)] = \sum_x P(x) f(x) \qquad E_{x\sim P}[f(x)] = \int f(x) p(x) dx$$

  - Weighted average
- if f(x)=x, expectation E(x) is called **mean**
- properties:

$$E[a \times f(x)] = a \times E[f(x)]$$

$$E[f(x) + g(x)] = E[f(x)] + E[g(x)]$$

- e.g. Bernoulli(ϕ), mean = ϕ

Variance

- The **variance** gives a measure of how much the values of a function of a random variable x vary as we sample different values of x from its probability distribution:

$$Var[f(x)] = E[(f(x) - E[f(x)])^2]$$

  - a measure of how concentrated the distribution is around the expectation
- Standard deviation = square root of variance
- usually, we use f(x)=x,
- Property:

$$Var[a \times f(x)] = a^2 \times Var[f(x)]$$

- e.g., Bernoulli(ϕ), variance = ϕ(1- ϕ)

Covariance

- The **covariance** gives some sense of how much two values are linearly related to each other, as well as the scale of these variables:

$$\operatorname{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])\left(g(y) - \mathbb{E}\left[g(y)\right]\right)\right]$$

$$Cov[X, Y] = E[X \times Y] - E[X] \times E[Y]$$

$$Cov[X, Y] = 0 \text{ if X, Y are independent}$$

  - High covariance values denote that both values change very much and far from the respective mean at the same time
  - correlation normalize the contribution of each variable to measure only how much the variables are related
  - Two variables with nonzero covariance are dependent

- Covariance matrix of an n-D r.v. is an nxn matrix

$$\operatorname{Cov}(\mathbf{x})_{i,j} = \operatorname{Cov}(\mathrm{x}_i, \mathrm{x}_j)$$

  - Diagonals of covariance matrix is variance

$$\operatorname{Cov}(\mathrm{x}_i, \mathrm{x}_i) = \operatorname{Var}(\mathrm{x}_i)$$

Bernoulli and Multinoulli distribution

- The **Bernoulli** distribution is a distribution over a single binary random variable

$$P(\mathrm{x}=1)=\phi$$

$$P(\mathrm{x}=0)=1-\phi$$

$$P(\mathrm{x}=x)=\phi^{x}(1-\phi)^{1-x}$$

$$E[x]=\phi$$

$$Var[x]=\phi(1-\phi)$$

- The **multinoulli** or **categorical distribution** is a distribution over a single discrete variable with k different states, where k is finite

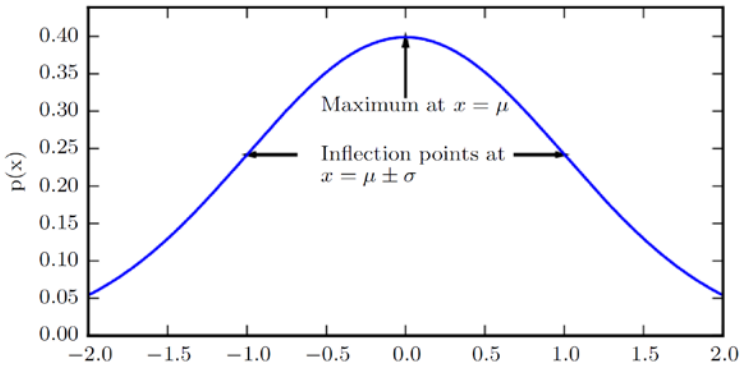
$$P(x)=\begin{cases}\phi_1 & \text{if } x=1\\ \vdots & \vdots\\ \phi_{k-1} & \text{if } x=k-1\\ 1-\sum_{i=1}^{k-1}\phi_i & \text{if } x=k\end{cases}$$

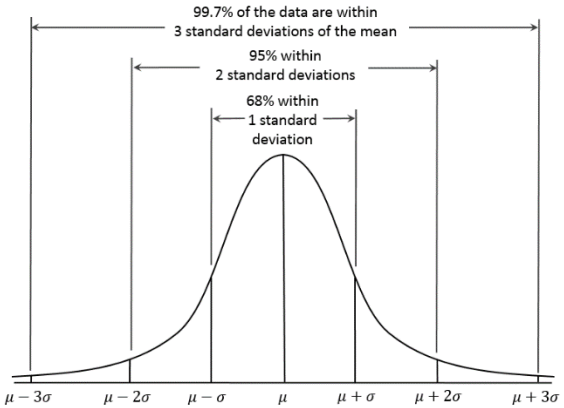
- **Multinomial distribution** (n,k) denotes how many times each of the k categories are visited when n samples are drawn from the distribution

Gaussian distribution

- Gaussian (or normal) distribution

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





known as 68-95-99.7 (empirical) rule, or 3-sigma rule

- normalized Gaussian $X \sim N(\mu,\sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$
- Multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

  - covariance matrix is often diagonal (called **isotropic** Gaussian), or even simpler of a scalar times an identity matrix

- **Central limiting theorem**: sum of many independent r.v. is approximated normally distributed

Empirical distribution

- all of the mass in a probability distribution clusters around a single point by defining a PDF using the **Dirac delta function**, δ(x):

$$p(x) = \delta(x - \mu) \triangleq \begin{cases} 1 & \text{if } x = \mu \\ 0 & \text{otherwise} \end{cases}$$

- **Empirical distribution** puts probability mass $1/m$ on each of the m points $x^{(1)}, \ldots, x^{(m)}$ forming a given dataset or collection of samples

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} \delta(\boldsymbol{x} - \boldsymbol{x}^{(i)})$$

  - Empirical distribution formed from a dataset of training examples specifies the distribution from which we sample when we train a model on this dataset
  - Empirical distribution is the probability density that maximizes the likelihood of the training data

Mixture distribution

- Made up of several component distributions
  - On each trial, the choice of which component distribution should generate the sample is determined by sampling a component identity from a multinoulli distribution

$$P(x) = \sum_i P(c = i) \times P(x \mid c = i),$$

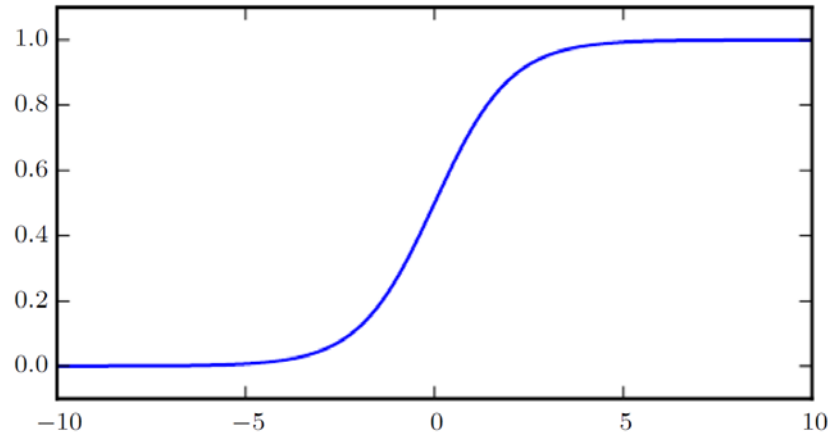
$P(c)$ is the multinoulli distribution over component identities

  - e.g., empirical distribution is a mixture distribution with one Dirac component distribution for each training sample
- Gaussian mixture model is when component distributions P(**x**|c) are Gaussian
  - Each component has its separately parameterized mean vector and covariance
  - could have constraints, e.g. covariance matrices of all components are identical, and diagonal or isotropic
  - Parameters of Gaussian mixture specify the **prior probability *P(c=i)***, expressing the model's belief about c ***before*** it has observed x
    - **P(c|x)** is a **posterior probability** because it is computed after the observation of x

Popular functions

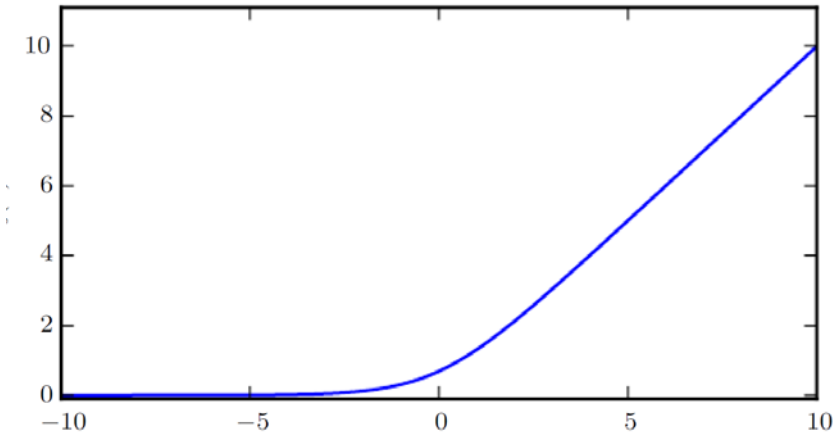
- Logistic sigmoid : $$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

  - commonly used to produce the ϕ parameter of Bernoulli(ϕ):



- Softplus: $$\zeta(x) = \log\left(1 + \exp(x)\right)$$

  - Softended version of $x^+ = \max(0, x)$



Bayes' Rule

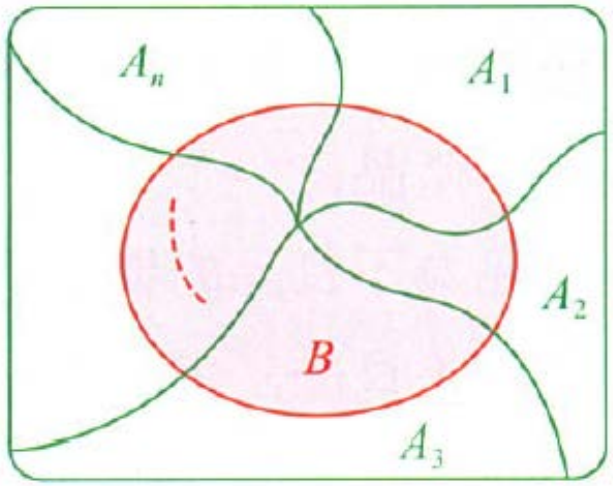
- we know P(y | x) and need to know P (x | y), if we also know P (x)

$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}.$$

- we do not need to know P(y), because it can be computed as

$$P(\mathrm{y}) = \sum_x P(\mathrm{y} \mid x)P(x)$$

- a sample space is divided into k disjoint subspace A_k



$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)}$$

random vector

- put multiple random variables into a vector $\mathbf{X} = [X_1 \quad X_2 \quad \cdots \quad X_n]^T$
- covariance matrix

$$
\begin{aligned}
\Sigma &= \begin{bmatrix} Cov[X_1, X_1] & \cdots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \cdots & Cov[X_n, X_n] \end{bmatrix} \\
&= \begin{bmatrix} E[X_1^2] - E[X_1]E[X_1] & \cdots & E[X_1X_n] - E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_nX_1] - E[X_n]E[X_1] & \cdots & E[X_n^2] - E[X_n]E[X_n] \end{bmatrix} \\
&= \begin{bmatrix} E[X_1^2] & \cdots & E[X_1X_n] \\ \vdots & \ddots & \vdots \\ E[X_nX_1] & \cdots & E[X_n^2] \end{bmatrix} - \begin{bmatrix} E[X_1]E[X_1] & \cdots & E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n]E[X_1] & \cdots & E[X_n]E[X_n] \end{bmatrix} \\
&= E[XX^T] - E[X]E[X]^T = \ldots = E[(X - E[X])(X - E[X])^T].
\end{aligned}
$$

- e.g. multi-variate Gaussian distribution

$$
f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)
$$

Information theory

- quantify how much information is present in a signal
  - an unlikely event has occurred is more informative than learning that a likely event has occurred
  - Likely events should have low information content
    - in the extreme case, events guaranteed to happen should have no information content
  - Less likely events should have higher information content
  - Independent events should have additive information
    - e.g., finding out that a tossed coin has come up as heads twice should convey twice as much information as finding out that a tossed coin has come up as heads once
- self-information of an event x = x, $I(x) = -\log P(x).$
- **Shannon entropy H(x)**: quantify the amount of uncertainty in an entire probability distribution

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]$$

Shannon entropy for binary r.v. with p

- entropy: (p-1)*log(1-p) – p*log(p)
  - high entropy when p=1/2, i.e., uniform binary r.v.
  - low entropy when the r.v. is close to deterministic, i.e., p is close to 1 or 0

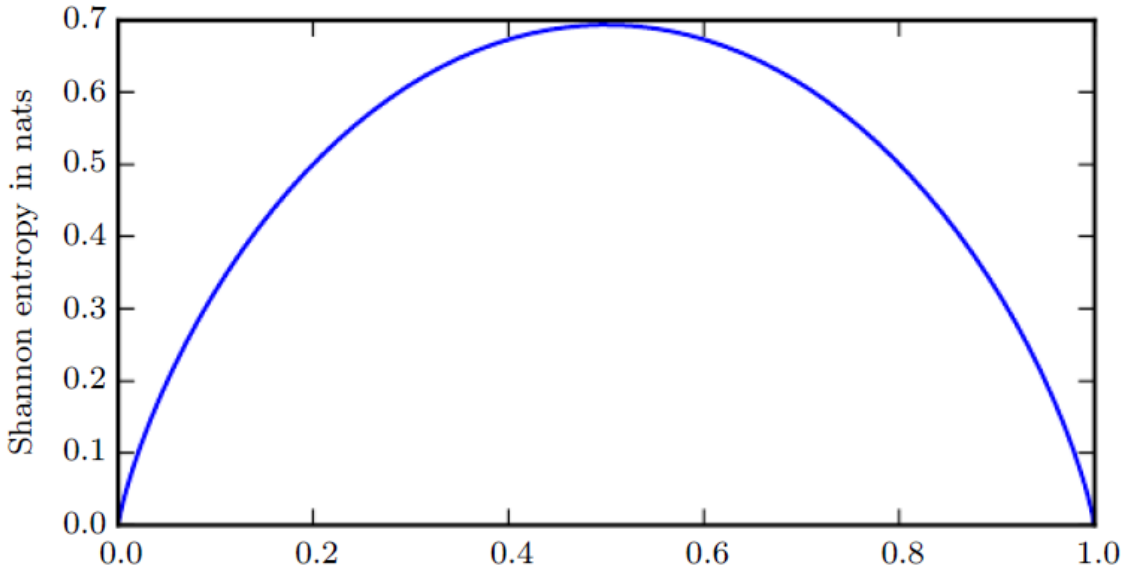


Figure 3.5: This plot shows how distributions that are closer to deterministic have low Shannon entropy while distributions that are close to uniform have high Shannon entropy. On the horizontal axis, we plot p, the probability of a binary random variable being equal to 1. The entropy is given by $(p-1)\log(1-p) - p\log p$. When p is near 0, the distribution is nearly deterministic, because the random variable is nearly always 0. When p is near 1, the distribution is nearly deterministic, because the random variable is nearly always 1. When $p = 0.5$, the entropy is maximal, because the distribution is uniform over the two outcomes.

cross entropy

- If we have two separate probability distributions P (x) and Q (x) over the same random variable x, we can measure how different these two distributions are using the **Kullback-Leibler (KL) divergence**:

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right]$$

  - P(x) is the distribution we want to approximate using another distribution Q(x)
  - KL divergence could be viewed as sort of distance between two distributions
  - e.g., in object classification, P(x) is the one-hot label and Q(x) is the computed
  - not symmetric, i.e., $D_{KL}(P||Q)$!= $D_{KL}(Q||P)$

- **cross entropy**: $H(P,Q) = H(P) + D_{KL}(P||Q)$
  - Minimizing the cross-entropy H(P,Q) with respect to Q is equivalent to minimizing the KL divergence, because Q does not participate in the omitted term

$$H(P,Q) = H(P) + D_{KL}(P\|Q) = -E_{x\sim P(x)}[\log Q(x)]$$

estimators

- Let a set of *m* i.i.d. data points: $\{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$
- the estimator is a function of the *m* data points

$$\hat{\theta}_m = g(x^{(1)}, x^{(2)}, \cdots, x^{(m)})$$

- e.g., *m* data points from Bernouli distribution with mean θ

$$\hat{\theta}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

- e.g., m data points from normal distribution $p(x^{(i)}) = N(x^{(i)}; \mu, \sigma^2)$

$$\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} \text{ : sample mean}$$

$$\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \hat{\mu}_m)^2 \text{ : sample variance}$$

bias of estimator

- bias of an estimator is defined as

$$\text{bias}(\hat{\theta}_m) = E(\hat{\theta}_m) - \theta$$

- an estimator is unbiased if

$$\text{bias}(\hat{\theta}_m) = E(\hat{\theta}_m) - \theta = 0$$

- an estimator is **asymptotically unbiased** (**consistency**) if

$$\text{bias}(\hat{\theta}_m) = \lim_{m \to \infty} E(\hat{\theta}_m) = \theta$$

bias of some estimators

- estimator for Bernouli distribution $\hat{\theta}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$

$$\text{bias}(\hat{\theta}_m) = E(\hat{\theta}_m) - \theta = E(\frac{1}{m}\sum_{i=1}^{m} x^{(i)}) - \theta$$

$$= \frac{1}{m}\sum_{i=1}^{m} \underbrace{E(x^{(i)})}_{\theta\times 1+(1-\theta)\times 0} - \theta = 0$$

- estimator for normal distribution

$$\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} \qquad E(\hat{\mu}_m) = \mu \Rightarrow \text{ unbias estimator}$$

$$\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \hat{\mu}_m)^2 \qquad E(\hat{\sigma}_m^2) = \frac{m-1}{m}\sigma^2 \Rightarrow \text{biased estimator}$$

Maximum Likelihood (ML) Estimation

- Consider a set of *m* examples $X = \{x^{(1)}, \ldots, x^{(m)}\}$ drawn independently from the true but unknown data generating distribution $p_{data}(x)$.
- Let $p_{model}(x; \theta)$ be a parametric family of probability distributions
- The maximum likelihood estimator for θ is then defined as

$$\theta_{ML} = \arg\max_{\theta} p_{\text{model}}(X;\theta) = \arg\max_{\theta} \prod_{i=1}^{m} p_{\text{model}}(x^{(i)};\theta)$$

- Take log, product is transformed into sum

$$\theta_{ML} = \arg\max_{\theta} p_{\text{model}}(X;\theta) = \arg\max_{\theta} \sum_{i=1}^{m} \log p_{\text{model}}(x^{(i)};\theta)$$

- Divided by m, expressed as an expectation w.r.t. empirical distribution $\hat{p}_{data}$ defined by the training data

$$\theta_{ML} = \arg\max_{\theta} E_{x\sim\hat{p}_{data}} [\log p_{\text{model}}(x;\theta)]$$

  - MLE minimizes the dissimilarity (KL divergence) between the empirical distribution $\hat{p}_{data}$, defined by the training set and the model distribution

$$D_{KL}(\hat{p}_{data} \| p_{\text{model}}) = E_{x\sim\hat{p}_{data}} [\log \hat{p}_{data}(x) - \log p_{\text{model}}(x)]$$

MLE example: normal distribution

- *n* samples from i.i.d. of normal distribution

$$f(x_1, x_2, \cdots, x_n \mid \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i \mid \mu, \sigma^2) = (\frac{1}{2\pi\sigma^2})^{n/2} \exp(\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2})$$

$$\log(L(\mu, \sigma)) = \log[f(x_1, x_2, \cdots, x_n \mid \mu, \sigma^2)] = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

- Let derivative of log-likelihood w.r.t. μ be zero, MLE for parameter μ is

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- let derivative of log-likelihood w.r.t. σ be zero, MLE for parameter σ is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Machine Learning example 1: Batch Normalization

$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$ [compute mini-batch mean]

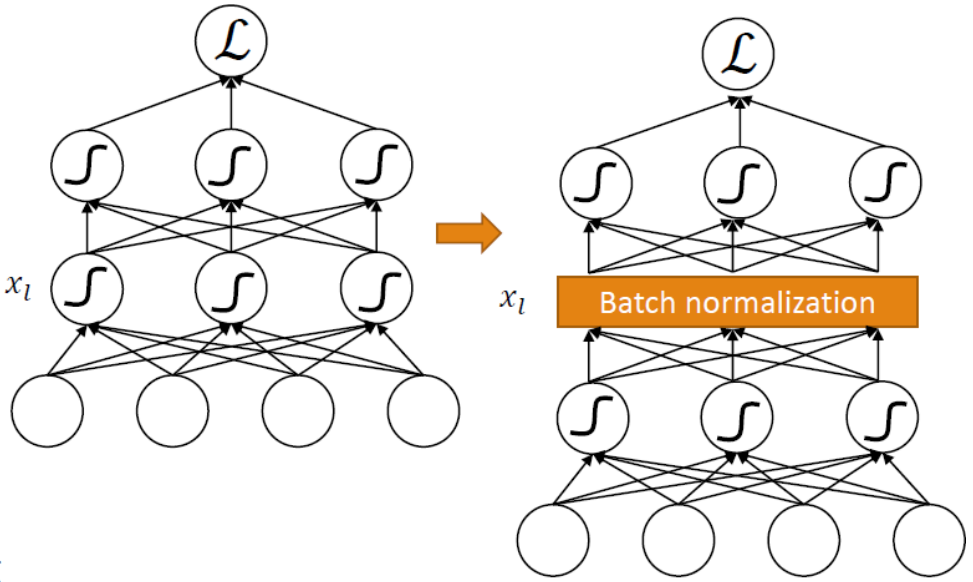
$\sigma_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2$ [compute mini-batch variance]

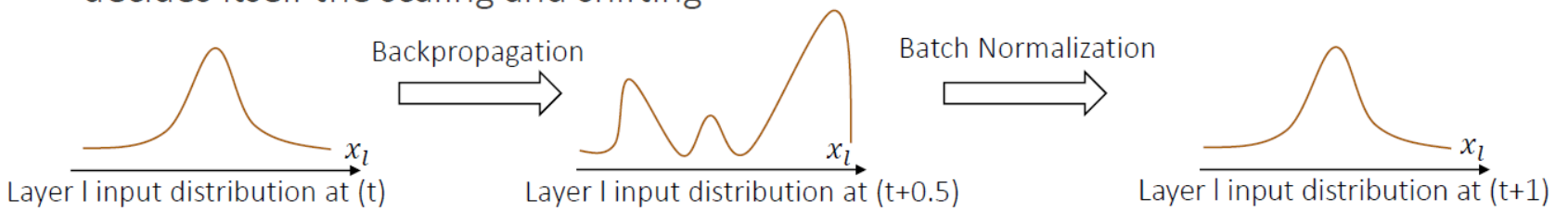
$\widehat{x_i} \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \varepsilon}}$ [normalize input]

$\widehat{y_i} \leftarrow \gamma x_i + \beta$ [scale and shift input]

Trainable parameters

- Weights change → the distribution of the layer inputs changes per round
- Normalize the layer inputs with batch normalization
  - Roughly speaking, normalize x_l to $N(0, 1)$, then rescale
  - Rescaling is so that the model decides itself the scaling and shifting



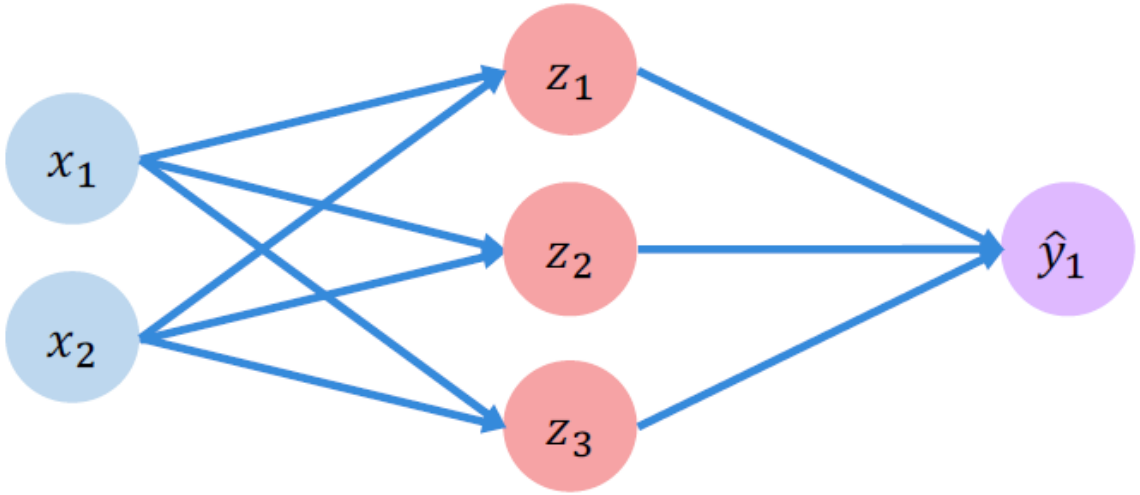


Machine Learning example 2

- input with two features: x_1 (# of attended lectures), x_2(# of study hours)

*The **empirical loss** measures the total loss over our entire dataset*

$$\boldsymbol{x} = \begin{bmatrix} 4, & 5 \\ 2, & 1 \\ 5, & 8 \\ \vdots & \vdots \end{bmatrix}$$



$f(x)$ y

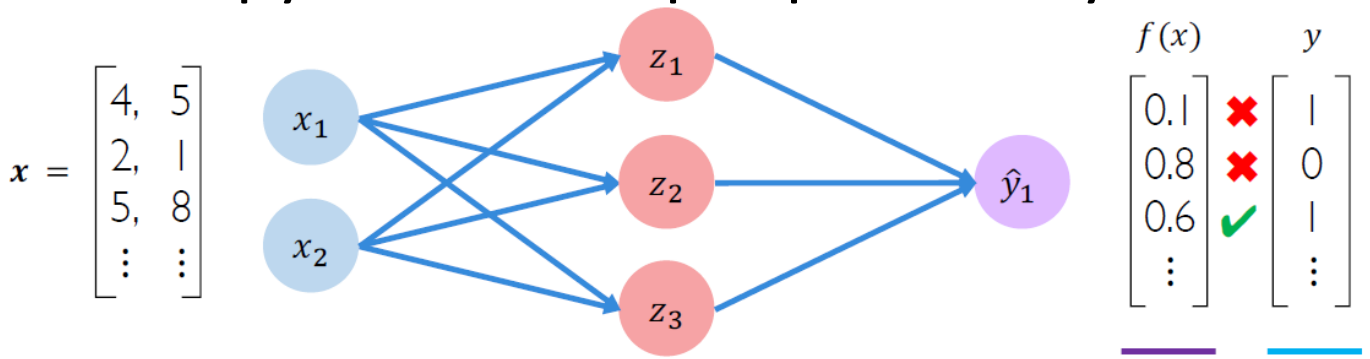
$$\begin{bmatrix} 0.1 \\ 0.8 \\ 0.6 \\ \vdots \end{bmatrix} \quad \begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \end{bmatrix}$$

Also known as:
- Objective function
- Cost function
- Empirical Risk

$$\boldsymbol{J(\theta)} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\left(\underbrace{f\left(x^{(i)};\theta\right)}_{\text{Predicted}}, \underbrace{y^{(i)}}_{\text{Actual}}\right)$$

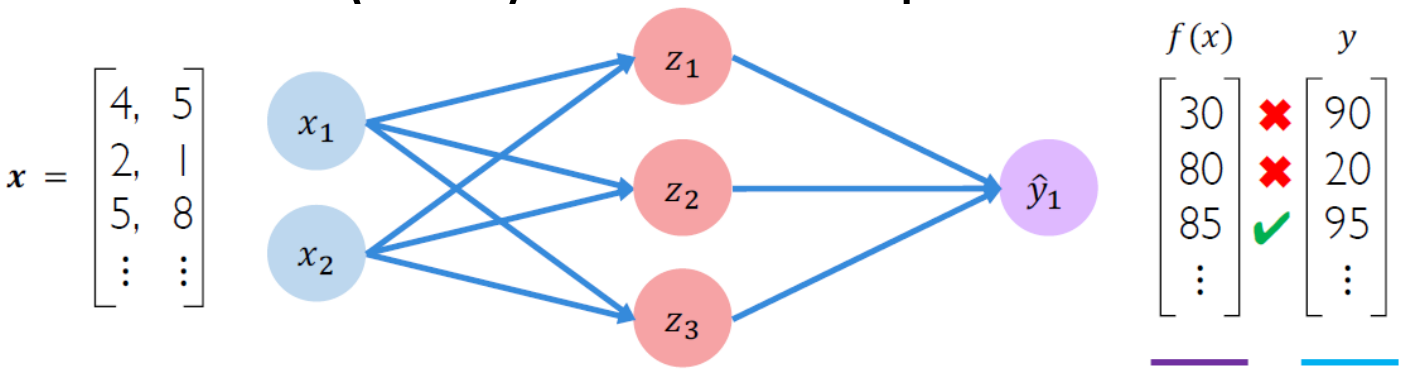
Loss functions

- binary cross entropy loss for output probability between 0 and 1



$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{y^{(i)}}_{\text{Actual}} \log\Big(\underbrace{f\big(x^{(i)};\theta\big)}_{\text{Predicted}}\Big) + (1-\underbrace{y^{(i)}}_{\text{Actual}}) \log\Big(1-\underbrace{f\big(x^{(i)};\theta\big)}_{\text{Predicted}}\Big)$$

- mean squared error (MSE) loss for output continuous numbers



$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \Big(\underbrace{y^{(i)}}_{\text{Actual}} - \underbrace{f\big(x^{(i)};\theta\big)}_{\text{Predicted}}\Big)^2$$

Example 3: object classification

Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W)$$

$$P(Y = k|X = x_i) = \frac{e^{s_k}}{\sum_j e^{s_j}}$$

Softmax Function

$$L_i = -\log P(Y = y_i|X = x_i)$$

	Unnormalized log-probabilities / logits	exp	Probabilities must be >= 0	normalize	Probabilities must sum to 1
cat	**3.2**		**24.5**		**0.13**
car	5.1		164.0		0.87
frog	-1.7		0.18		0.00
	unnormalized probabilities				probabilities

→ L_i = -log(0.13)
= **2.04**

Maximum Likelihood Estimation
Choose probabilities to maximize the likelihood of the observed data
(See CS 229 for details)

Cross Entropy

Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W)$$

$$P(Y = k|X = x_i) = \frac{e^{s_k}}{\sum_j e^{s_j}}$$

Softmax Function

Probabilities must be >= 0

Probabilities must sum to 1

$$L_i = -\log P(Y = y_i|X = x_i)$$

	Unnormalized log-probabilities / logits	exp	unnormalized probabilities	normalize	probabilities	compare	Correct probs
cat	3.2		24.5		0.13		1.00
car	5.1		164.0		0.87		0.00
frog	-1.7		0.18		0.00		0.00

Kullback–Leibler divergence

$$D_{KL}(P||Q) = \sum_y P(y) \log \frac{P(y)}{Q(y)}$$

Cross Entropy

$$H(P, Q) = H(p) + D_{KL}(P||Q)$$