Chapter 2 - Probability

Statement: "the probability that a coin will land heads is 0.5"

Frequentist interpretation: if we flip the coin many times, we expect it to land head half the time, i.e., probabilities represent long run frequencies of events
Bayesian interpretation: on the next toss, the coin is equally likely to land heads or tails. In this view, probability is used to quantify uncertainity about something.
In many cases, we do not have long run frequencies of events, or we cannot perform an experiment N number of times to come up with probability values - for example, whether the polar ice cap will melt in 2030? - we may have to compute this probability using other factors.

- Conditional probability: $P(A|B) = P(A,B)/P(B)$
- Joint probability: $P(A,B) = P(A|B) \times P(B)$
- Joint = conditional times marginal
  This is called the **product rule of probability**.
- Marginal distribution of A (by marginalizing over B): $P(A) = \Sigma_b P(A, B = b) = \Sigma_b P(A|B = b) \times P(B = b)$ (**sum rule** or **rule of total probability**)
- Applying the product rule over a joint distribution of multiple RVs, we get the **chain rule**:
  $P(X_{1:D}) = P(X_1) \times P(X_2|X_1) \times P(X_3|X_1, X_2)....P(X_D|X_{1:D-1})$
- **Bayes Rule**:
  $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} == \frac{P(B|A) \times P(A)}{\Sigma_a P(B|A=a) \times P(A=a)}$ (denominator expanded using the sum rule)

Generative Classifier:

We compute $p(y = c|\mathbf{x}, \theta)$ using the bayes rule after first computing the **class conditional probability density** $p(\mathbf{x}|y = c, \theta)$ and multiplying it with the class prior $p(y = c)$

Discriminative Classifier:

Directly fits $p(y = c|\mathbf{x}, \theta)$

Unconditional independence or marginal independence:

$X \perp Y \iff P(X,Y) = P(X) \times P(Y)$
Joint = product of marginals

Conditional independence:

Given a third variable Z, X and Y are conditionally independent iff :
$X \perp Y|Z \iff P(X,Y|Z) = P(X|Z) \times P(Y|Z)$

For discrete RVs, we have probability mass functions. For continuous RVs, we first define cummulative distribution function $F(X = a) = P(X \leq a)$ and then take its derivative to obtain the probability density function $p(X)$

Common Discrete Distributions

1. Binomial: two outcomes; p is the probability of success
2. Bernoulli: Repeat binomial experiment n times
3. Multinoulli/Categorical distribution: k possible outcomes, 1-of-K encoding (rolling a dice, softmax across vocabulary)
4. Multinomial: Repeat multinoulli experiment n times
5. Poisson: a model for **counts** of rare events, e.g. number of accidents on a street. $Poi(x|\lambda) = e^{-\lambda} * \frac{\lambda^x}{x!}$, where λ is known as the rate parameter.

Common Continuous Distributions

1. Gaussian

2. Student t distribution - one problem with the Gaussian distribution is that it is sensitive to outliers since the log probability only decays quadratically with the distance from the centre. Student t is more robust to outliers. For larger values of the `degrees of freedom` parameter, Student t approaches Gaussian.

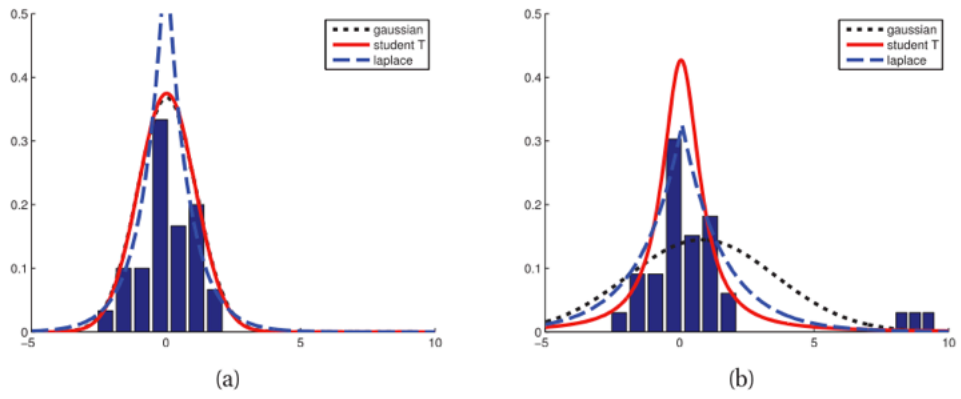
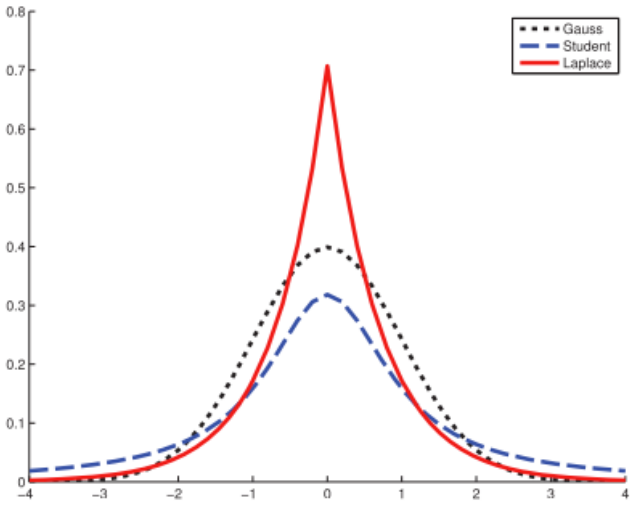


Figure 2.8 Illustration of the effect of outliers on fitting Gaussian, Student and Laplace distributions. (a) No outliers (the Gaussian and Student curves are on top of each other). (b) With outliers. We see that the Gaussian is more affected by outliers than the Student and Laplace distributions. Based on Figure 2.16 of (Bishop 2006a). Figure generated by `robustDemo`.

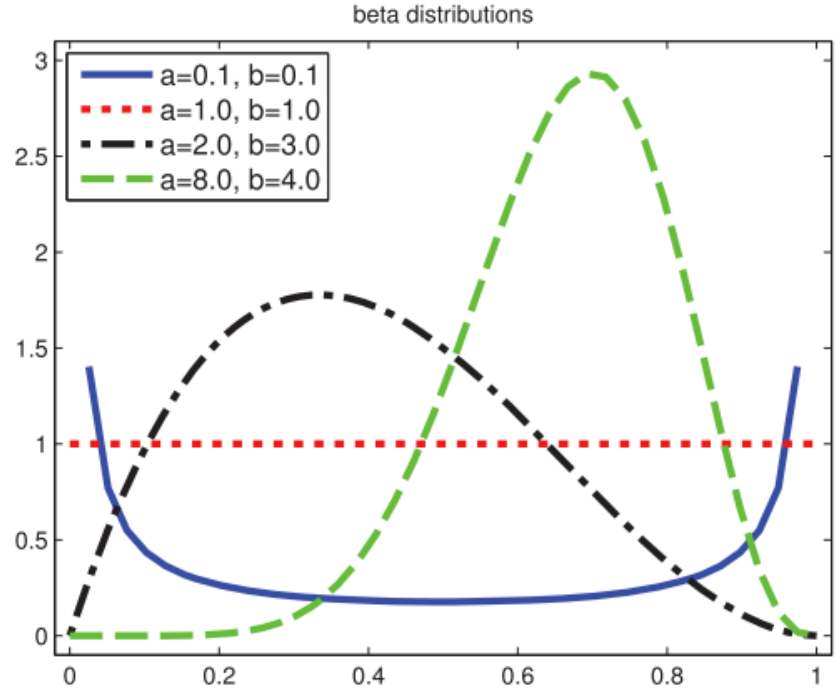
3. Laplace Distribution - essentially a double sided exponential distribution



4. Gamma distribution: two parameters `a=shape` and `b=rate`. Special cases:

- Exponential distribution (a=1, b=λ)
- Erlang distribution (a=2, b=λ)
- Chi-squared distribution: a RV defined as the sum of squared Gaussian RVs

5. Beta distribution:



Joint Probability Distributions

- So far we focussed on modelling univariate probability distributions. A more challenging problem is modelling joint probability distributions on multiple related RVs.
- If X and Y are independent, then cov[X,Y]=0
- However, the reverse is NOT true. If cov[X,Y]=0, it does not mean that X and Y are independent. This is because cov/corr corresponds to only linear dependence, but X and Y may have non-linear dependence.
- Multi-variate Gaussian distribution

- **Dirichlet Distribution** : multi-variate generalization of the beta distribution. It has a support over the probability simplex.

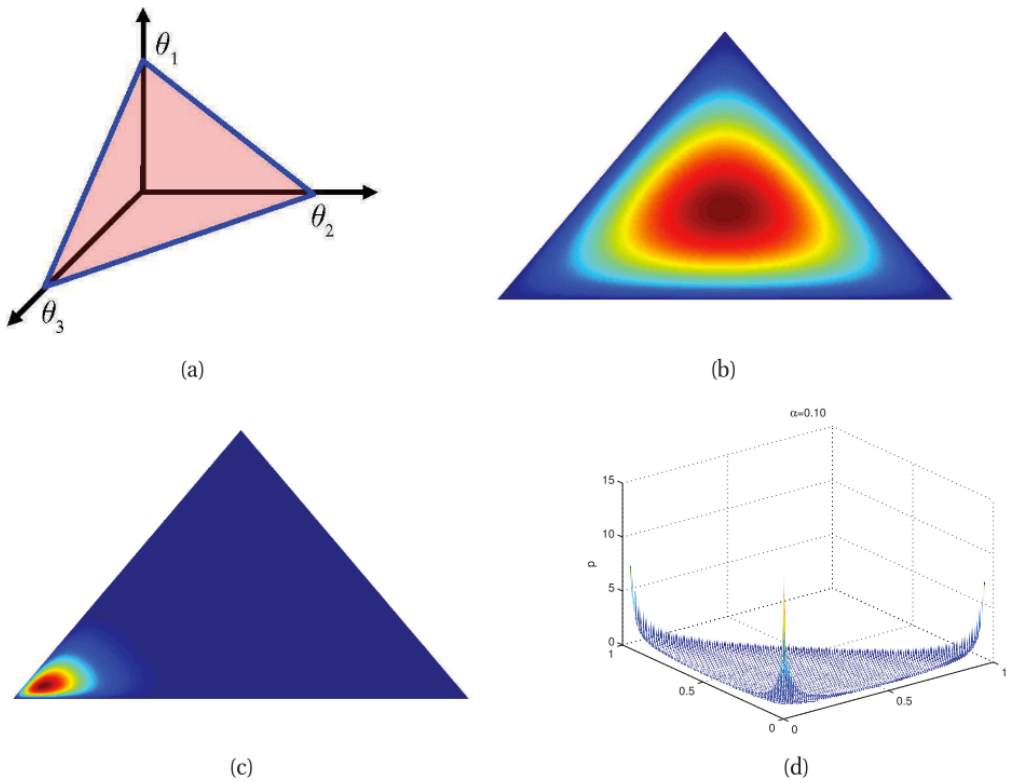


Figure 2.14 (a) The Dirichlet distribution when $K = 3$ defines a distribution over the simplex, which can be represented by the triangular surface. Points on this surface satisfy $0 \leq \theta_k \leq 1$ and $\sum_{k=1}^{3} \theta_k = 1$. (b) Plot of the Dirichlet density when $\boldsymbol{\alpha} = (2, 2, 2)$. (c) $\boldsymbol{\alpha} = (20, 2, 2)$. Figure generated by `visDirichletGui`, by Jonathan Huang. (d) $\boldsymbol{\alpha} = (0.1, 0.1, 0.1)$. (The comb-like structure on the edges is a plotting artifact.) Figure generated by `dirichlet3dPlot`.

Transformations of RVs:

Linear transformations:
$\mathrm{E}[a^T \mathbf{x} + b] = a^T \mu + b$
$\mathrm{V}ar[a^T \mathbf{x} + b] = a^T \boldsymbol{\Sigma} a$ (Σ is the covariance matrix)

General Transformation (univariate case):
$p_y(y) = p_x(x) \left| \frac{dx}{dy} \right|$
Suppose $X = U(0, 1)$ and $Y = X^2$, then $p_x(x) = 1$ in the range (0,1) and $p_y(y) = \frac{1}{2\sqrt{y}}$

In the multivariate case to map $\mathbf{y} = f(\mathbf{x})$, one need to compute the Jacobian.

$$\mathbf{J}_{\mathbf{x} \to \mathbf{y}} \triangleq \frac{\partial (y_1, \ldots, y_n)}{\partial (x_1, \ldots, x_n)} \triangleq \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & amp; \cdots & amp; \frac{\partial y_1}{\partial x_n} \\ \vdots & amp; \ddots & amp; \vdots \\ \frac{\partial y_n}{\partial x_1} & amp; \cdots & amp; \frac{\partial y_n}{\partial x_n} \end{pmatrix}$$

|det J| measures how much a unit cube changes in volume when we apply f
If f is an **invertible mapping**, we can define the pdf of the transformed variables using the Jacobian of the inverse mapping $\mathbf{y} \to \mathbf{x}$:

$$p_y(\mathbf{y}) = p_x(\mathbf{x}) \left| \det \left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right) \right| = p_x(\mathbf{x}) \, |\det \mathbf{J}_{\mathbf{y} \to \mathbf{x}}|$$

Example: Consider transforming a density from Cartesian coordinates $x = (x1, x2)$ to polar coordinates $y = (r, \theta)$,where $x1 = rcos\theta$ and $x2 = rsin\theta$.

Step 1: compute $\mathbf{J_{y \to x}}$

Step 2: find $|\det \mathbf{J}|$

Finally,

$$p_{\mathbf{y}}(\mathbf{y})amp; = p_{\mathbf{x}}(\mathbf{x})|\det \mathbf{J}|$$

$$p_{r,\theta}(r,\theta)amp; = p_{x_1,x_2}(x_1,x_2)\, r = p_{x_1,x_2}(r\cos\theta, r\sin\theta)|r|$$

Central Limit Theorem

- N random variables with pdfs (not necessarily Gaussian) $p(x_i)$ each with its own mean and variance.
- Assume that each variable is independent and identically distributed.
- As N increases, the distribution of the RV formed by the sum $S_N = \Sigma_i X_i$ will approach a Gaussian distribution.

Monte Carlo approximation

- In general, computing the distribution of a function of an rv using the change of variables formula can be difficult (especially in the multi-variate case since we need to compute the det of the Jacobian).
- A simple alternative is to generate S samples from the original distribution by Markov chain Monte Carlo (MCMC) sampling. Given these samples, we can approximate the distribution of $f(X)$ by using the empirical distribution of $\{f(x_s)\}_{s=1}^{S}$
- The accuracy of an MC approximation increases with sample size.

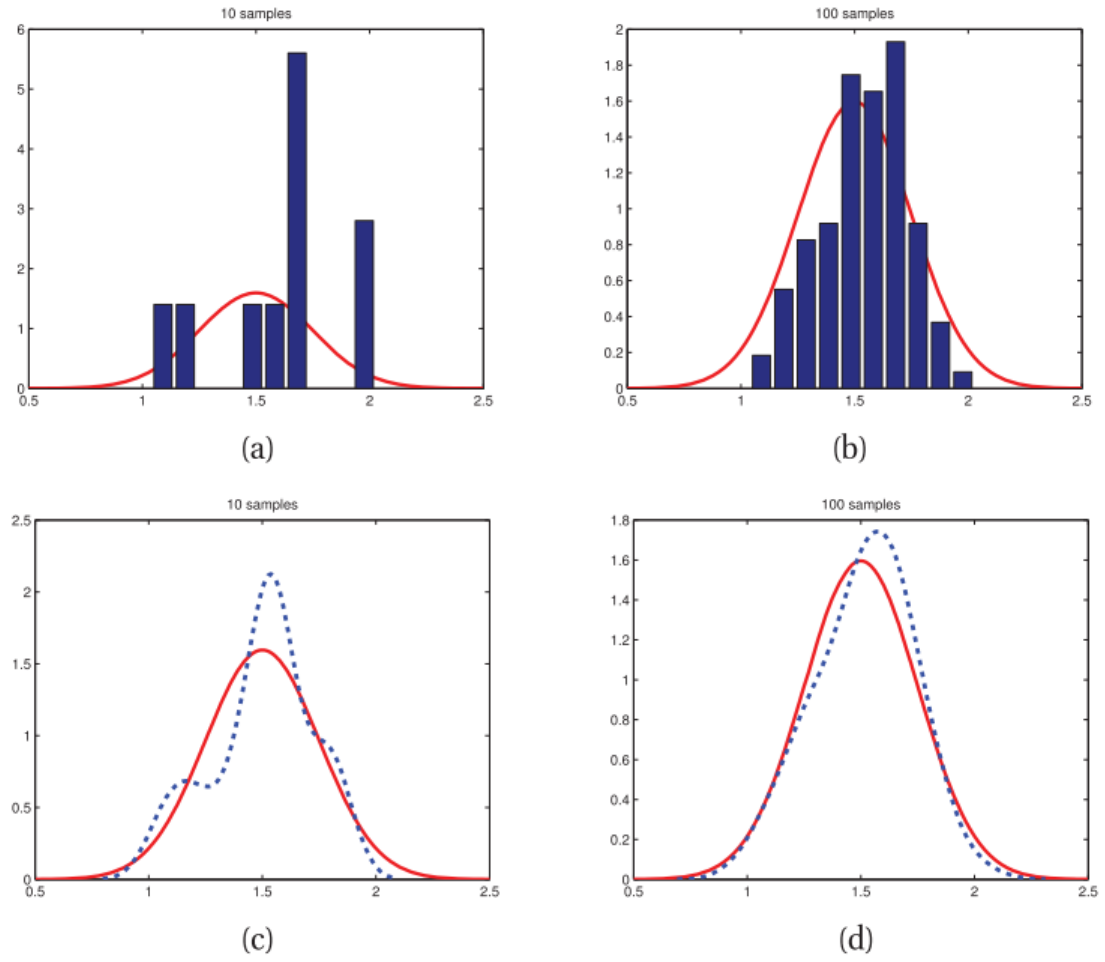


Figure 2.20 10 and 100 samples from a Gaussian distribution, $\mathcal{N}(\mu = 1.5, \sigma^2 = 0.25)$. Solid red line is true pdf. Top line: histogram of samples. Bottom line: kernel density estimate derived from samples in dotted blue, solid red line is true pdf. Based on Figure 4.1 of (Hoff 2009). Figure generated by `mcAccuracyDemo`.

- Suppose we compute the mean of $f(X)$ as $\hat{\mu}$, we can report the standard error as a measure of how confident we are on the estimate of the mean of the transformed distribution. $\hat{\sigma}^2 = \frac{1}{S}\sum_{s=1}^{S}(f(x_s) - \hat{\mu})^2$
- Then, from CLT, we have the probability that the estimate is within 95% confidence interval of the original μ as $P\{\mu - 1.96\frac{\hat{\sigma}}{\sqrt{S}} \leq \hat{\mu} \leq \mu + 1.96\frac{\hat{\sigma}}{\sqrt{S}}\} \approx 0.95$
- The term $\sqrt{\frac{\hat{\sigma}^2}{S}}$ is called the **standard error**, and is an estimate of our uncertainty about our estimate of μ.
- If we want to report an answer which is accurate to within $\pm\epsilon$ with probability at least 95%, we need to use a number of samples S which satisfies $1.96\sqrt{\hat{\sigma}^2/S} \leq \epsilon$. We can approximate the 1.96 factor by 2, yielding $S \geq \frac{4\hat{\sigma}^2}{\epsilon^2}$.

Information Theory

- Information theory is concerned with representing data in a compact fashion (a task known as data compression or source coding).

Entropy

- a measure of uncertainity. For a discrete RV with K states,

$$\mathrm{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k)$$

- If log base 2, the measurement unit for entropy is called **bits** (short for binary digits). If we use log base e, the units are called **nats**.
- The discrete distribution with maximum entropy is the uniform distribution.
- Laplace's principle of insufficient reason, which argues in favor of using uniform distributions when there are no other reasons to favor one distribution over another.
- Minimum entropy (of zero) when we have a dirac delta function at one of the K classes.

KL Divergence

- One way to measure the dissimilarity of two probability distributions, p and q, is known as the Kullback-Leibler divergence (KL divergence) or relative entropy.

$$\mathrm{KL}(p\|q) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$$

- We can re-write this as:

$$\mathrm{KL}(p\|q) = \sum_k p_k \log p_k - \sum_k p_k \log q_k = -\mathrm{H}(p) + \mathrm{H}(p,q)$$

- where $\mathrm{H}(p,q)$ is called **cross-entropy**.
- One can show that cross entropy is the average number of bits needed to encode data coming from a source with distribution p when we use model q to define our codebook.
- Hence the "regular" entropy $\mathrm{H}(p) = \mathrm{H}(p,p)$, is the expected number of bits **if we use the true model**, so the KL divergence is the difference between these.
- In other words, the KL divergence is the **average number of extra bits** needed to encode the data, due to the fact that we used distribution q to encode the data instead of the true distribution p.
- The "extra number of bits" interpretation should make it clear that $\mathrm{KL}(p\|q) \geq 0$, and that the KL is only equal to zero iff $q = p$ (this can be formally proved using Jensen's inequality).
- $\mathrm{KL}(p\|q)$: This is called **forward** KL. However, in Bayesian inference and VAEs, **reverse** KL is widely used ($\mathrm{KL}(q\|p)$).
- The KL divergence is not a distance, since it is asymmetric. One symmetric version of the KL divergence is the **Jensen-Shannon divergence**.

$$D_{JS}(p\|q) = \frac{1}{2}D_{KL}\left(p\|\frac{p+q}{2}\right) + \frac{1}{2}D_{KL}\left(q\|\frac{p+q}{2}\right)$$

Mutual Information

- We know that correlation coefficient only measures linear dependence.
- An alternative measure to know the dependence between two RVs X and Y is MI: Determine how similar

(in terms of KL div) the joint distribution $p(X, Y)$ is to the factored distribution $p(X)p(Y)$

- $$\mathrm{I}(X;Y) \triangleq \mathrm{KL}(p(X,Y) \| p(X)p(Y)) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$
- One can re-write the above as $\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X|Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X)$
- where $\mathrm{H}(Y|X)$ is the **conditional entropy** defined as $\mathrm{H}(Y|X) = \sum_x p(x)\mathrm{H}(Y|X = x)$. Thus, one can interpret MI between X and Y as the **reduction in uncertainty** about X after observing Y.
- **Pointwise mutual information** or PMI between two events (not random variables) x and y:
- $$\mathrm{PMI}(x,y) \triangleq \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}$$
- This measures the discrepancy between these events occuring together compared to what would be expected by chance. Clearly, MI is simply the expected value of PMI.
- The above formulae are defined for discrete events/RVs. For continuous RVs, it is common to first discretize or quantize them (Unfortunately, the number of bins used, and the location of the bin boundaries, can have a significant effect on the results.).
- An alternative metric for continuous RVs, in order to measure even non-linear relationships is the MIC (**maximal information coefficient** in the range [0,1]).