EC-332 Machine Learning

Nazar Khan

Department of Computer Science
University of the Punjab

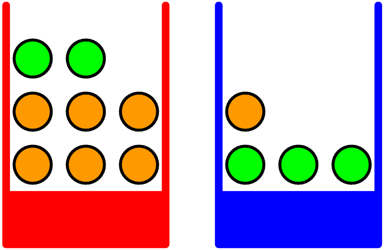
Probability

Probability Theory

- **Uncertainty** is a key concept in pattern recognition.
- Uncertainty arises due to
  - Noise on measurements.
  - Finite size of data sets.
- Uncertainty can be **quantified** via probability theory.

Probability

- $P(\text{event})$ is fraction of times event occurs out of total number of trials.
- $P = \lim_{N\to\infty} \frac{\#successes}{N}$.



$$P(B = b) = 0.6$$

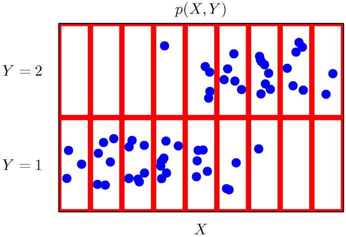
$$P(B = r) = 0.4$$

$$p(apple) = p(F = a) = ?$$

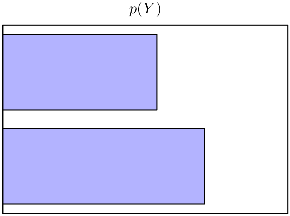
$$p(\text{blue box given that apple was selected}) = p(B = b|F = a) = ?$$

Terminology

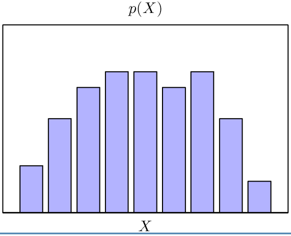
Joint Probability



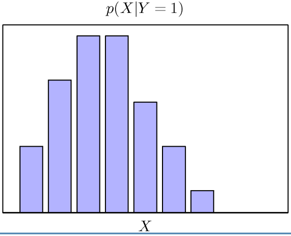
Marginal Probability



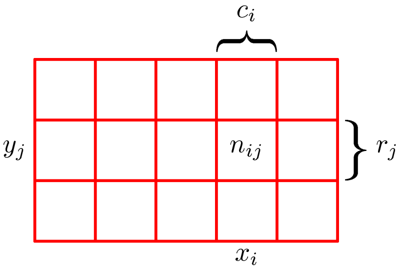
Marginal Probability



Conditional Probability



Elementary rules of probability



Elementary rules of probability

- Sum rule: $p(X) = \sum_Y p(X, Y)$
- Product rule: $p(X, Y) = p(Y|X)p(X)$

These two simple rules form the basis of *all* the probabilistic machinery that will be used in this course.

- The sum and product rules can be combined to write

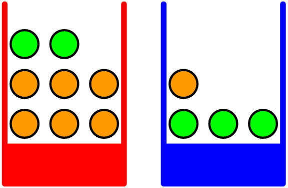
$$p(X) = \sum_Y p(X|Y)p(Y)$$

- A fancy name for this is **Theorem of Total Probability**.
- Since $p(X, Y) = p(Y, X)$, we can use the product rule to write another very simple rule

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

- Fancy name is **Bayes' Theorem**.
- Plays an *important role* in machine learning.

Terminology

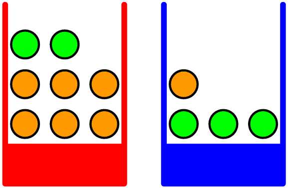


$$P(B = r) = 0.4$$

$$P(B = b) = 0.6$$

- If you don't know which fruit was selected, and I ask you which box was selected, what will your answer be?
  - The box with greater probability of being selected.
  - Blue box because $P(B = b) = 0.6$.
  - This probability is called the **prior probability**.
  - Prior because the data has not been observed yet.

Terminology



$$P(B = r) = 0.4$$
$$P(B = b) = 0.6$$

- Which box was chosen given that the selected fruit was orange?
  - The box with greater $p(B|F = o)$ (via Bayes' theorem).
  - Red box
  - This is called the **posterior probability**.
  - Posterior because the data has been observed.

Independence

- If random variable X is **independent** of random variable Y, then

$$P(X = x|Y = y) = P(X = x)$$

  for all values x and y.
- Then, by the product rule

$$P(X, Y) = P(X|Y)P(Y) = P(X)P(Y)$$

- If joint $p(X = x, Y = y)$ equals the product of marginals $p(X = x)p(Y = y)$ *for all values x and y*, then random variables X and Y are **independent**.
- Intuitively, if Y is independent of X, then knowing X does not change the chances of Y and vice versa.
- Example: if fraction of apples and oranges is same in both boxes, then knowing which box was selected does not change the chance of selecting an apple.

Probability density
Continuous

- So far, our set of events was discrete.
- Probability can also be defined for continuous variables via

$$\text{Prob}(x \in (a, b)) = \int_a^b p(x)dx$$

- *Probability density function* $p(x)$
  - is always non-negative, and
  - integrates to 1.

Caution: Probability density is not the same as probability. Density can be greater than 1.

Probability density
Continuous

- Sum rule: $p(x) = \int p(x, y) dy$.
- Product rule: $p(x, y) = p(y|x)p(x)$
- Probability density can also be defined for a multivariate random variable $\mathbf{x} = (x_1, \ldots, x_D)$.

$$p(\mathbf{x}) \geq 0$$
$$\int_{\mathbf{x}} p(\mathbf{x}) d\mathbf{x} = \int_{x_D} \cdots \int_{x_1} p(x_1, \ldots, x_D) dx_1 \ldots dx_D = 1$$

Expectation

- Expectation is a weighted average of a function.
- Weights are given by $p(x)$.

$$\mathbb{E}[f] = \sum_x p(x) f(x) \qquad \longleftarrow \text{For discrete } x$$

$$\mathbb{E}[f] = \int_x p(x) f(x) dx \qquad \longleftarrow \text{For continuous } x$$

- When data is finite, expectation $\approx$ ordinary average. Approximation becomes exact as $N \rightarrow \infty$ (*Law of large numbers*).

Expectation

- Expectation of a function of several variables

$$\mathbb{E}[f(x,y)] = \sum_{x,y} p(x,y) f(x,y)$$

- Expectation with respect to one variable

$$\mathbb{E}_x[f(x,y)] = \sum_{x} p(x) f(x,y) \qquad \text{(function of } y\text{)}$$

- *Conditional expectation*

$$\mathbb{E}_{x|y}[f] = \sum_{x} p(x|y) f(x)$$

Variance

- *Variance* measures variability of a random variable around its mean.

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

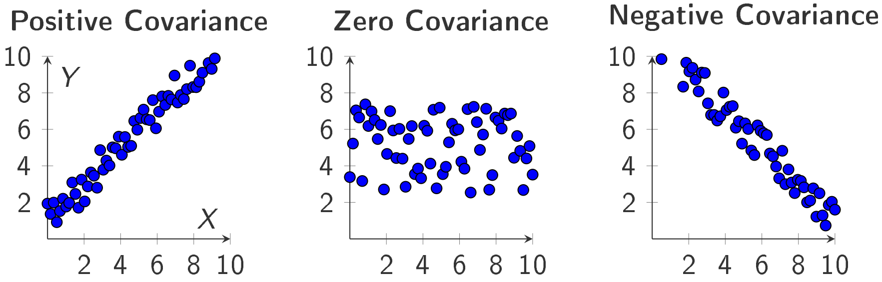
- On average, how far does a random variable stay from its mean?

Covariance
Univariate

- For 2 univariate random variables occuring in pairs (x, y), **covariance** expresses how much x and y vary together.

$$cov[x, y] = \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}]$$
$$= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

- For independent random variables x and y, $cov[x, y] = 0$. Why?



Covariance
Multivariate

- For multivariate random variables $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{y} \in \mathbb{R}^K$, $cov[\mathbf{x}, \mathbf{y}]$ is a $D \times K$ matrix.
- Expresses how each element of $\mathbf{x}$ varies with each element of $\mathbf{y}$.

$$
\begin{aligned}
cov[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y} - \mathbb{E}[\mathbf{y}]\}^T\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\mathbf{x}\mathbf{y}^T\right] - \mathbb{E}[\mathbf{x}]\,\mathbb{E}[\mathbf{y}]^T \\
&= \begin{bmatrix} cov[x_1, y_1] & cov[x_1, y_2] & \cdots & cov[x_1, y_K] \\ cov[x_2, y_1] & cov[x_2, y_2] & \cdots & cov[x_2, y_K] \\ \vdots & \vdots & \ddots & \vdots \\ cov[x_D, y_1] & cov[x_D, y_2] & \cdots & cov[x_D, y_K] \end{bmatrix}
\end{aligned} \tag{1}
$$

Covariance
Multivariate

- Covariance of multivariate $\mathbf{x}$ with itself can be written as $cov[\mathbf{x}] \equiv cov[\mathbf{x}, \mathbf{x}]$.
- $cov[\mathbf{x}]$ expresses how each element of $\mathbf{x}$ varies with every other element.

$$cov[\mathbf{x}] = \begin{bmatrix} var[x_1] & cov[x_1, x_2] & \cdots & cov[x_1, x_D] \\ cov[x_2, x_1] & var[x_2] & \cdots & cov[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ cov[x_D, x_1] & cov[x_D, x_2] & \cdots & var[x_D] \end{bmatrix} \quad (2)$$

Bayesian View of Probability

- So far we have considered probability as the *frequency of random, repeatable events*.
- What if the events are not repeatable?
  - Was the moon once a planet?
  - Did the dinosaurs become extinct because of a meteor?
  - Will the ice on the North Pole melt by the year 2100?
- For non-repeatable, yet uncertain events, we have the **Bayesian view** of probability.

Bayesian View of Probability

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

- Measures the uncertainty in model $\mathbf{w}$ after observing the data $\mathcal{D}$.
- This uncertainty is measured via conditional $p(\mathcal{D}|\mathbf{w})$ and prior $p(\mathbf{w})$.
- Treated as a function of $\mathbf{w}$, the conditional probability $p(\mathcal{D}|\mathbf{w})$ is also called the **likelihood function**.
- Expresses how likely the observed data is for any given model $\mathbf{w}$.