Expectation

Definition

The expectation of a term $t(\mathbf{x})$ with respect to the random variable $\mathbf{x}$ is defined as

$$\mathbf{E}_{\mathbf{x}}\langle t(\mathbf{x})\rangle := \int_{\mathcal{X}} t(\mathbf{x})d\Pr(\mathbf{x}) = \int_{\mathcal{X}} t(\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

The last equation is valid if a density exists.

Example: Uniform Distribution

Assume the uniform distribution on $[0,10]$. What is the expected value of $t(\mathbf{x}) = \mathbf{x}^2$?

$$\mathbf{E}_{\mathbf{x}}\langle t(\mathbf{x})\rangle = \int_{[0,10]} t(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \int_{[0,10]} \mathbf{x}^2\frac{1}{10}d\mathbf{x} = 33\frac{1}{3}$$

Example: Roulette

What is the expected loss in roulette when we bet on a number, say j (we win 36\$:1\$ if the number is hit and 0\$:1\$ otherwise)?

$$\mathbf{E}_{\mathbf{x}}\langle t(\mathbf{x})\rangle = \sum_{i=1,i\neq j}^{37} -1\$\cdot\frac{1}{37} + 35\$\cdot\frac{1}{37} = -\frac{1}{37}\$$$

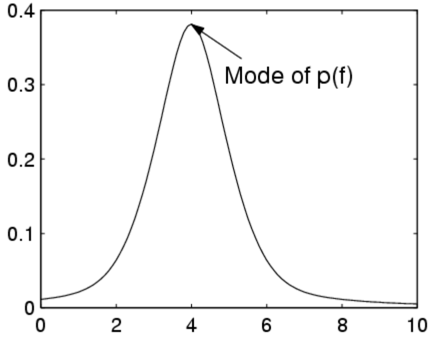
Mean and Mode

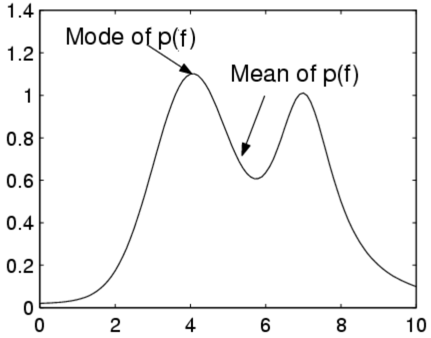
Mean

It is the expected value of the random variable $\mathbf{x}$ itself, i.e. $\mu := \mathbf{E}_{\mathbf{x}}[\mathbf{x}]$

Mode

It is the largest value of the density $p(\mathbf{x})$. This corresponds to the most frequently observed values of $\mathbf{x}$. Note that mode and mean in general not coincide.





Variance

Definition

It is the amount of variation of the random variable. We can obtain this by first standardizing $\mathbf{x}$ such that it has zero mean and subsequently computing the second order moment of the new random variable.

$$\sigma^2 := \mathbf{E}_{\mathbf{x}}\left[(\mathbf{x}-\mathbf{E}_{\mathbf{x}}[\mathbf{x}])^2\right] = \mathbf{E}_{\mathbf{x}}\left[\mathbf{x}^2 - 2\mathbf{x}\mathbf{E}_{\mathbf{x}}[\mathbf{x}] + (\mathbf{E}_{\mathbf{x}}[\mathbf{x}])^2\right] = \mathbf{E}_{\mathbf{x}}\mathbf{x}^2 - (\mathbf{E}_{\mathbf{x}}[\mathbf{x}])^2$$

Normalization

A useful way of preprocessing data is to rescale them to zero mean and unit variance, i.e. $\sigma^2 = 1$ (we call σ the standard devitation). This is obtained by $\mathbf{x} \to \frac{\mathbf{x}-\mu}{\sigma}$

Tails of Distributions

Note that the variance need not always exist. Long-tailed distributions can be killers for insurance companies (e.g. distributions of earthquakes, storms, etc. — strong ones are very unlikely but still may happen).

Example: $\Pr(i) = \frac{1}{\zeta(3)i^3}$, where i is the damage incurred by i.

Chebyshev's Inequality

Chebyshev's Inequality

For any random variable $\mathbf{x}$ we can bound deviations of $\mathbf{x}$ from its mean $\mathbf{E}_{\mathbf{x}}[\mathbf{x}]$ by

$$\Pr(|\mathbf{x}-\mu| > C) \leq \frac{\sigma^2}{C^2}$$

Proof

All we have to do is compute an upper bound for σ and backtrack.

$$\begin{aligned}\sigma^2 &= \int |\mathbf{x}-\mu|^2 dP(\mathbf{x}) = \int_{|\mathbf{x}-\mu|>C} |\mathbf{x}-\mu|^2 dP(\mathbf{x}) + \int_{|\mathbf{x}-\mu|\geq C} |\mathbf{x}-\mu|^2 dP(\mathbf{x}) \\ &\geq \int_{|\mathbf{x}-\mu|>C} |\mathbf{x}-\mu|^2 dP(\mathbf{x}) \geq C^2\Pr(|\mathbf{x}-\mu| > C)\end{aligned}$$

Applications

In engineering quite often we get information about some measurement and about the variance of the quantity. But usually we do not know the distribution.

Still we want to make statements about the probability that some devices will exceed their specifications.

Normal Distribution

The Formula

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

Mean

The mean of $p(x)$ is μ. We can see this by showing that $p(\mu + \xi) = p(\mu - \xi)$. This follows immediately from $((\mu + \xi) - \mu)^2 = \xi^2 = ((\mu - \xi) - \mu)^2$.

Variance

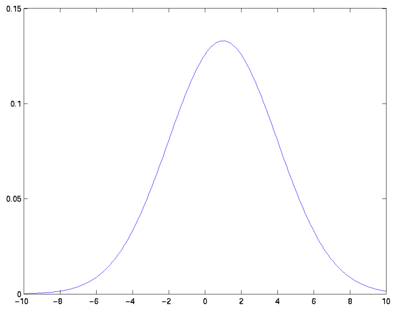
The variance of $p(x)$ is σ^2. We show this by proving that

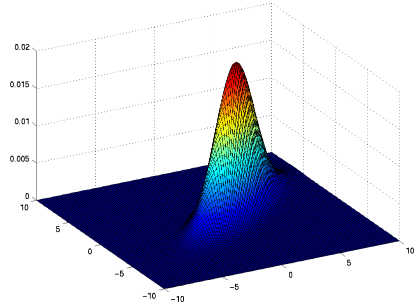
$$\begin{aligned}
\mathrm{Var}x &= \int_{\mathbb{R}} p(x)(x-\mu)^2 dx = \int_{\mathbb{R}} p(\mu+\xi)\xi^2 d\xi \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{-\frac{\xi^2}{2\sigma^2}} \xi^2 d\xi \\
&= \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{\xi^2}{2}} \xi^2 d\xi = \sigma^2.
\end{aligned}$$

Pictures of Normal Distributions

Normal Distribution in $\mathbb{R}$: Mean 1, Variance 3

Normal Distribution in $\mathbb{R}^2$: Mean $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$, Variance $\begin{bmatrix} 6 & 4 \\ 4 & 4 \end{bmatrix}$





Variance and Covariance

Covariance

First off, we have to define the variance of a multivariate distribution. We set this to be

$$\mathrm{Cov}\mathbf{x} = \mathbf{E}\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top\right]$$

This means that we compute a **matrix** rather than just a single number, telling us the **correlation** between random variables. In particular

$$(\mathrm{Cov}\mathbf{x})_{ij} = \mathbf{E}\left[(x_i - \mu_i)(x_i - \mu_i)\right]$$

Correlated Variables

If two random variables are completely correlated, we have

$$(\mathrm{Cov}\mathbf{x})_{ij} = \sqrt{(\mathrm{Cov}\mathbf{x})_{ii}(\mathrm{Cov}\mathbf{x})_{jj}}$$

For anticorrelated variables, the sign is reversed.

Uncorrelated Variables

A necessary condition for uncorrelated variables is that their covariance vanishes.

Multivariate Normal Distribution

The Formula

It is essentially a product of several univariate normal distributions.

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^m \det\Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1} (\mathbf{x}-\mu)\right)$$

Here instead of the scalar σ^{-2} we have a positive definite matrix $\Sigma \in \mathbb{R}^{m \times m}$, and the mean becomes a vector $\mu \in \mathbb{R}^m$.

Mean Obviously this is μ (we can check that by symmetry)

Variance

Now eigenvalues and eigenvectors come in handy. We decompose $\Sigma = O^\top \Lambda O$, where O is an orthogonal matrix and Λ is diagonal.

$$\begin{aligned}
\mathrm{Var}\mathbf{x} &= \frac{1}{\sqrt{(2\pi)^m \det\Sigma}} \int_{\mathbb{R}} \mathbf{x}\mathbf{x}^\top \exp\left(-\frac{1}{2}\mathbf{x}^\top \Sigma^{-1} \mathbf{x}\right) d\mathbf{x} \\
O(\mathrm{Var}\mathbf{x})O^\top &= \frac{1}{\sqrt{(2\pi)^m \det\Sigma}} \int_{\mathbb{R}} O\mathbf{x}\mathbf{x}^\top O^\top \exp\left(-\frac{1}{2}\mathbf{x}^\top O^\top O \Sigma^{-1} O^\top O\mathbf{x}\right) d\mathbf{x}
\end{aligned}$$

Multivariate Normal Distribution, Part II

Variance

We use $\det\Sigma = \det O\Sigma O^\top = \prod_i \lambda_i$, where λ_i are the eigenvalues of Σ.

Next use the variable transform $\mathbf{x} \to O\mathbf{x}$ for the integraction and we obtain

$$\begin{aligned} O(\mathrm{Var}\mathbf{x})O^\top &= \frac{1}{\sqrt{(2\pi)^m \prod_{i=1}^m \lambda_i}} \int_{\mathbb{R}} \mathbf{x}\mathbf{x}^\top \exp\left(-\frac{1}{2}\mathbf{x}^\top \Lambda^{-1}\mathbf{x}\right) d\mathbf{x} \\ &= \Lambda \end{aligned}$$

This is so since Λ is diagonal, hence all off diagonal terms of $O(\mathrm{Var}\mathbf{x})O^\top$ vanish. As for the diagonal terms we get the one-dimensional normal distribution.

Therefore $\mathrm{Var}\mathbf{x} = O^\top \Lambda O = \Sigma$. Hence we proved that the multivariate Gaussian has variance Σ.

Conditioning on Normal Distribution

Multivariate Normal Distributions

Assume that $(\mathbf{x}, x) \in \mathbb{R}^{m+1}$ is distributed according to a normal distribution, given by zero mean ($\mu = 0$) and a covariance matrix $\begin{bmatrix} \Sigma & \boldsymbol{\sigma} \\ \boldsymbol{\sigma}^\top & s \end{bmatrix}$. Clearly, the distribution is nicely centered around 0.

Observing Variables

Now we observe $\mathbf{x}$. This allows us to compute $p(x|\mathbf{x})$ from $p(\mathbf{x}, x)$ via $p(\mathbf{x}|\mathbf{y}) = \dfrac{p(\mathbf{x}, \mathbf{y})}{\int_{\mathcal{X}} p(\mathbf{x}, \mathbf{y}) d\mathbf{x}}$ We have

$$\begin{bmatrix} \Sigma & \boldsymbol{\sigma} \\ \boldsymbol{\sigma}^\top & s \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma^{-1} - \chi^{-1}\left(\Sigma^{-1}\boldsymbol{\sigma}^\top\right)^\top \left(\Sigma^{-1}\boldsymbol{\sigma}^\top\right) & -\chi^{-1}\left(\Sigma^{-1}\boldsymbol{\sigma}^\top\right) \\ -\chi^{-1}\left(\Sigma^{-1}\boldsymbol{\sigma}^\top\right)^\top & \chi^{-1} \end{bmatrix}$$

where $\chi = s - \boldsymbol{\sigma}^\top \Sigma^{-1} \boldsymbol{\sigma}$.

Conditioning on Normal Distribution II

Useful Trick

We want to avoid carrying out the integral over $\mathbf{x}$. This can be done by realizing that the restriction of a normal distribution on a subset of variables is a normal distribution again. So, all we need to know are **mean** and **variance** with respect to x.

Variance

All that survives from the quadratic term in $\begin{bmatrix} \mathbf{x} \\ x \end{bmatrix}^\top \begin{bmatrix} \Sigma & \boldsymbol{\sigma} \\ \boldsymbol{\sigma}^\top & s \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x} \\ x \end{bmatrix}$ is χ^{-1}, so the variance of x is reduced to $\chi = s - \boldsymbol{\sigma}^\top \Sigma^{-1} \boldsymbol{\sigma}$ from s which we had without $\mathbf{x}$.

Mean

We need to make a quadratic extension and obtain that

$$p(x) = \frac{1}{\sqrt{2\pi\chi}} \exp\left(-\frac{1}{2\chi}(x - \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\sigma})^2\right)$$

Laplacian Distribution

Example: Decay of Atoms

The probability that a certain atom does not decay within 1s is p, hence the probability that it will decay within n seconds is $1 - p^n$.

The continuous version thereof is to assume that after time t the probability of decay is

$$P(\xi \leq T) = 1 - \exp(-\lambda T) = \int_0^T p(t) dt.$$

Laplacian Distribution

Consequently, $p(t)$ is given by $\lambda \exp(-\lambda T)$. This distribution is called the **Laplacian** distribution.

It is a particularly long-tailed distribution. We also find this in cases such as the discharge of capacitors, etc.

Laplacian Distribution, Part II

Mean

The mean of the distribution (expected lifetime of an atom) is given by (we use partial integration)

$$\mu = \int_0^\infty t\lambda \exp(-\lambda t) dt = \int_0^\infty \exp(-\lambda t) dt = \frac{1}{\lambda}.$$

Variance

For convenience, we use the symmetric Laplacian distribution $p(t) = \frac{\lambda}{2}\exp(-\lambda t)$ for $t \in \mathbb{R}$. This distribution has clearly zero mean. The variance is given by

$$\frac{\lambda}{2}\int_{\mathbb{R}} t^2 \exp(-\lambda t) dt = 2\int_{[0,\infty]} t\exp(-\lambda t) dt = \frac{2}{\lambda}\int_{[0,\infty]} \exp(-\lambda t) dt = \frac{2}{\lambda^2}$$

Likelihood

Likelihood of Data

For a density $p(\mathbf{x})$ depending on parameters θ, also denoted by $p(\mathbf{x}|\theta)$, we can determine the likelihood of an observation $\mathbf{x}_0$ by $p(\mathbf{x}_0|\theta)$

Likelihood of IID Sample

For a series of observations $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, drawn iid from $p(\mathbf{x}|\theta)$ we may write the likelihood by

$$p(X|\theta) = p(\mathbf{x}_1|\theta) \cdot \ldots \cdot p(\mathbf{x}_m|\theta)$$

Log-Likelihood

Quite often (for optimization purposes) it is convenient to take the logarithm of the likelihood and we obtain

$$\mathcal{L} = \log p(X|\theta) = \sum_{i=1}^m \log p(\mathbf{x}_i|\theta)$$

This allows us to find the most appropriate parameter θ, given the data, by maximizing $p(X|\theta)$, i.e. the plausibility of the data, given the parameter.

One Dimensional Case

Maximum Likelihood for Normal Distribution

We assume that some numbers, say $X := \{x_1, \ldots, x_m\}$ were generated by the normal distribution $p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$.

Find the value of μ that maximizes the likelihood that the data was generated by $p(X|\mu)$. The log likelihood is given by

$$\mathcal{L} = \log\left[(2\pi\sigma^2)^{-\frac{m}{2}} \prod_{i=1}^m \exp(-\frac{1}{2\sigma^2}(x_i-\mu)^2)\right] = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^m (x_i-\mu)^2$$

Mean and Normal Distribution

We minimize the negative log likelihood $-\mathcal{L}$ with respect to μ and obtain

$$\partial_\mu - \mathcal{L} = \frac{1}{\sigma^2}\sum_{i=1}^m (\mu - \xi_i) = 0 \text{ and therefore } \mu = \frac{1}{m}\sum_{i=1}^m \xi_i.$$

Law of Large Numbers

Why Gaussians are good for you If we have many independent errors, the net effect will be a single error with normal distribution.

Theorem

Denote by ξ_i random variables with variance $\sigma_i \leq \bar{\sigma}$ for some $\bar{\sigma}$ and with mean $\mu_i \leq \bar{\mu}$ for some μ, then the random variable $\xi := \frac{\sum_{i=1}^m \xi_i - \mu_i}{\sqrt{\sum_{i=1}^m \sigma_i^2}}$has zero mean and unit variance.

Furthermore for $m \to \infty$ the random variable ξ will be normally distributed.

Proof

Zero mean is simple. The fact that it will converge to a normal distribution is rather tricky. So we settle for the variance.

$$\mathrm{Var}\xi = \left[\sum_{i=1}^m \sigma_i^2\right]^{-1} \mathbf{E}\left[\sum_{i=1}^m (\xi_i - \mu_i)\right]^2 = \left[\sum_{i=1}^m \sigma_i^2\right]^{-1} \mathbf{E}\sum_{i,j=1}^m (\xi_i-\mu_i)(\xi_j-\mu_j) = 1.$$

Hoeffding's Bound

Sum of Random Variables

What happens if we have m random variables $\xi_i \in [0,1]$ and we average

$$\xi := \frac{1}{m}\sum_{i=1}^{m} \xi_i.$$

Will the fluctuations of ξ_i cancel out and will ξ be concentrated around its mean?

Hoeffing's Theorem

For any $\varepsilon > 0$ the probability of large deviations of ξ from $\mathbf{E}[\xi]$ is bounded by

$$\Pr\left(|\xi - \mathbf{E}[\xi]| \geq \varepsilon\right) \leq 2\exp\left(-2\varepsilon^2 m\right)$$

This means that things get exponentially better, the more random variables we average over. The practical use in our case is that things will get **exponentially better with the number of observations** in the training set.

Curse of Dimensionality

Bellman's Observation

If the number of dimensions of the input increases, estimators will get exponentially worse with the dimensionality.

Basic Idea

The distance between points increases with the dimensionality. For two randomly chosen points $\mathbf{x}, \mathbf{x}' \in [0,1]^m$ according to the uniform distribution we have

$$\mathbf{E}\left[\|\mathbf{x} - \mathbf{x}'\|^2\right] = \mathbf{E}\left[\sum_{i=1}^{m}(\mathbf{x}_i - \mathbf{x}_i')^2\right] = m\int_{-1}^{1}|1-x|x^2 dx = \frac{m}{6}.$$

Density

If 100 points are distributed on $[0,10]$, we have roughly 10 points per unit interval, for $[0,10]^2$, it's rougly onek and in $\mathbb{R}^m$ we have 10^{2-m} points.

Rule of Thumb

For each dimension we need roughly 10 observations.