Probabilistic Properties – Recap.

- The **sum, product** and **Bayes** rules.
- Expectation, variance, covariance:
  $var[f] = E[(f(x) - E[f(x)])^2] = E[f(x)^2] - E[f(x)]^2$
  $cov[x, y] = E[(x - E[x])(y - E[y])]$

Frequentist vs. Bayesian:

Frequentist: probabilities as frequencies of random repeatable events (the outcome of infinite trials). E.g. $p(heads) = 0.1$ yields that a coin tossed 1000 times will result in "heads" 100 times.

Bayesian: probabilities as uncertainties. If the chance of the next coin toss is 10% for heads, then $p(heads) = 0.1$.

Bayesian Probability

Let w be a vector of numbers we want to estimate, given data (observations) D, then:

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

- Posterior probability: $p(w|D)$ – given the knowledge D
- Likelihood: $p(D|w)$ (the generative model)
- Prior probability: $p(w)$ – no knowledge is given
- Normalization: $p(D)$

Let w be a vector of test scores, and D be observations of a set of people. We want to know from D what are the scores we need to get a degree. We assume $p(D)$ is uniform in the sense that we see any kind of samples. We want to pick the combination of scores w that maximizes the probability to get a degree.

$p(w)$ is the probability of seeing certain scores, and different w's can be plugged in.

So the Bayesian approach: try to maximize the **posterior**, which is proportional to the **prior X likelihood**.

- Frequentist – maximizing the likelihood: $argmax_w p(D|w)$
- Bayesian – maximizing the posterior: $argmax_w p(w|D)$

Example:

Let w be a random variable over in $\{0,1\}$ (say, fail and pass a coin toss), and let D be a training data: $D = \{1,1,1\}$. In that case:

Frequentist: $argmax_w p(D|w) = 1$, because this particular data got us 1 all 3 times.

Bayesian: $argmax_w p(w|D) = argmax_w \frac{p(D|w)p(w)}{p(D)} \underset{ignore\ p(D)}{=} \frac{1}{2}$ – which makes much more sense ($p(D|w) = 1$ but $p(w) = \frac{1}{2}$).

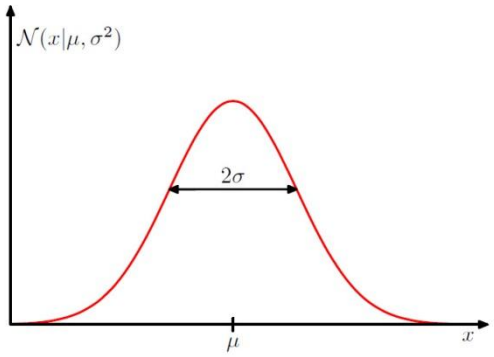
Probability Distribution

The most frequently used is **Gaussian**:

$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)0.5}\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$

Where:

- μ is the mean
- σ^2 is the variance (and precision: $\beta = \frac{1}{\sigma^2}$)



Observe that the exponent is negative (negation of product of squares), and so the maximum value will be when $x = \mu$.

Properties:

- $\mathcal{N}(x|\mu,\sigma^2) > 0$
- $\int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)dx = 1$

As said:

- $E[x] = \mu$
- Second moment $E[x^2] = \mu^2 + \sigma^2$
- $\Rightarrow var[x] = \sigma^2$

Central Limit Theorem

If we take the sum of random variables across many trials, the distribution of that sum values will become Gaussian.

Gaussian distribution

Consider $\mathbb{x} = [x_1, \dots, x_N]^T$ a vector of observations in which the data points are independently and identically distributed (i.i.d) Gaussian. Then:

$$p(x|\mu,\sigma^2) = \prod_{n=1}^{N}\mathcal{N}(x_n|\mu,\sigma^2)$$

Maximum-likelihood estimates of Gaussian distribution:

By the Frequentist approach we want to maximize the expression above. To do that we can take the derivative with respect to the variable we are interested in (either μ or σ^2). We will take the logarithm to get rid of the product:

$$\ln p(x|\mu,\sigma^2) = \ln\left(\prod_{n=1}^{N}\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - N\frac{1}{2}\ln\sigma^2 - N\frac{1}{2}\ln 2\pi$$

Now the derivatives to find the maximum:

$\underline{\mu}$: $(\ln p(x|\mu,\sigma^2))' = \frac{\partial p(x|\mu,\sigma^2)}{\partial \mu} = 0 \Leftrightarrow \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) = 0 \Leftrightarrow \boxed{\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n}$

So we just confirmed that μ is indeed the mean

*Note: logarithm doesn't change the trend, so finding a maximum of the log is equivalent to finding the maximum.

Now for σ^2:

$\underline{\sigma^2}$: $\frac{\partial p(x|\mu,\sigma^2)}{\partial \sigma} = 0 \Leftrightarrow \cdots \Leftrightarrow \boxed{\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2}$ which is what we expect – the mean of the squared difference of the mean.

In conclusion:

- $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$
- $\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$

Calculating expectations of those estimators:

$E[\mu_{ML}] = E\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] = \frac{1}{N}E[\sum_{n=1}^{N} x_n] = \frac{1}{N}\cdot N \cdot \mu = \mu \Rightarrow$ meaning, we get the true mean.

$$E[\sigma^2_{ML}] = E\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2\right] = \frac{1}{N}E\left[\sum_{n=1}^{N}(x_n - \mu_{ML})^2\right] = \frac{1}{N}\sum_{n=1}^{N}E[(x_n - \mu_{ML})^2] = *$$

$E[(x_n - \mu_{ML})^2] = E[x_n^2 - 2x_n\mu_{ML} + \mu^2_{ML}] = E[x_n^2] - 2E[x_n\mu_{ML}] + E[\mu^2_{ML}] = \cdots ?$

$$* = \cdots = \sigma^2 + \mu - \frac{1}{N^2}(N\sigma^2 - N^2\mu^2) = \boxed{\frac{N-1}{N}\sigma^2}$$

And we see that $E[\sigma^2_{ML}] \neq \sigma^2$, therefore it is **biased**

To make it unbiased we need to take $\frac{N}{N-1}\sigma^2_{ML}$.

Multivariate Gaussian Distribution

$$\mathcal{N}(\mathbb{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}}\cdot\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\mathbb{x}-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

Where μ is D-dimensional and Σ is the covariance $D \times D$ matrix (and $|\Sigma|$ is the determinant).

In the 1-D case we have the exponent $\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$

We can think of it as: $\exp\left[-\frac{1}{2}y^2\right]$ where $y = \frac{x-\mu}{\sigma}$

This makes it a $\mu = 0, \sigma = 1$ distribution which is the normal distribution.

In the multivariate case:

$y = U(x - \mu)$ where U is the matrix whose rows are the eigenvectors of the covariance matrix Σ.

$$p(y) = p(x)|J| = \prod_{j=1}^{D}\frac{1}{(2\pi\lambda_j)^{\frac{1}{2}}}\exp\left[-\frac{y_j^2}{2\lambda_j}\right]$$

Where λ_j are the eigenvalues of the covariance matrix Σ.

Multivariate ML Estimation

Let $\mathbb{X} = [\mathbb{x}_1, \dots, \mathbb{x}_N]^T$ be a data set with i.i.d. Gaussian observations $\mathbb{x}_n$:

$$\mathcal{N}(\mathbb{X}|\mu, \Sigma) = p(\mathbb{X}|\mu, \Sigma) = \prod_{n=1}^{N} \mathcal{N}(\mathbb{x}_n|\mu, \Sigma)$$

Then the log likelihood is calculated in the same way as for 1-D, and we get again:

- ML mean: $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{x}_n$
- ML covariance: $\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\mathbb{x}_n - \mu_{ML})\,(\mathbb{x}_n - \mu_{ML})^T$

(The slides show the formula of $\ln\big(p(\mathbb{X}|\mu, \Sigma)\big)$)

And we can also show that:

- $E[\mu_{ML}] = \mu$
- $E[\Sigma_{ML}] = \frac{N-1}{N}\Sigma$

And the unbiased estimator will be: $\tilde{\Sigma} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbb{x}_n - \mu_{ML})\,(\mathbb{x}_n - \mu_{ML})^T$

Sequential maximum likelihood

Denote the ML estimator of the mean based on N observations as $\mu_{ML}^{(N)}$:

$$\mu_{ML}^{(N)} = \frac{1}{N}\sum \mathbb{x}_n = \frac{1}{N}\mathbb{x}_N + \frac{1}{N}\sum_{}^{N-1} \mathbb{x}_n = \frac{1}{N}\mathbb{x}_N + \frac{N-1}{N}\mu_{ML}^{(N-1)} = \boxed{\mu_{ML}^{(N-1)} + \frac{1}{N}\big(\mathbb{x}_N - \mu_{ML}^{(N-1)}\big)}$$

So what it means is that given a new data point, we can easily correct our estimator by adding the difference of the current estimate of the mean: $\frac{1}{N}\big(\mathbb{x}_N - \mu_{ML}^{(N-1)}\big)$, making this value sequential.

Bayesian Inference of Gaussian

Let x be a random variable and X a set of N observations, and let σ^2 be given. We want to infer μ:

Likelihood:

$$p(X|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right]$$

We need a **conjugate** prior which should be a Gaussian

Prior:

$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$

Our goal is to find what the Gaussian would be given a Gaussian likelihood and prior.

The product:

We're not doing a straight-forward multiplication. Instead, consider the following D-dimensional generalization:

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(X-\mu) = -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu + const$$

And we are looking at the term that is linear to the mean and the variance: $x^T\Sigma^{-1}\mu$

In our case:

$$p(\mu|X) \propto p(X|\mu)p(\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right]\cdot\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left[-\frac{1}{2\sigma_0^2(\mu-\mu_0)^2}\right]\Rightarrow$$

Looking only at the exponent linear in μ:

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(-2x_n\mu)-\frac{1}{2\sigma_0^2}(-2\mu\mu_0)=\frac{1}{\sigma^2}\mu\sum_{n=1}^{N}x_n+\frac{1}{\sigma_0^2}\mu\mu_0=\frac{N}{\sigma^2}\mu_{ML}\mu+\frac{1}{\sigma_0^2}\mu\mu_0=$$

$$\mu\underbrace{\frac{\sigma_0^2N+\sigma^2}{\sigma^2\sigma_0^2}}_{1}\underbrace{\left(\frac{N\sigma_0^2}{\sigma_0^2N+\sigma^2}\mu_{ML}+\frac{\sigma^2}{\sigma_0^2N+\sigma^2}\mu_0\right)}_{2}$$

Where the (1) is $\frac{1}{\sigma_{MAP}^2}$ of μ and (2) is μ_{MAP}

From Bayes and by completing the square in the exponent:

MAP estimation:

Posterior:

$p(\mu|X)=\mathcal{N}(\mu|\mu_N,\sigma_N^2)$ where:

$$\mu_N=\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0+\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}$$

μ_N is a weighted sum of μ_0 and μ_{ML}, where the latter gets more weight as N increases.

And:

$$\frac{1}{\sigma_N^2}=\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}$$

That's computing the posterior probability from the likelihood and the prior.

The posterior of $N-1$ points can be viewed as the prior of observing the Nth data point:

$$p(\mu|X)\propto\left[p(\mu)\prod_{n=1}^{N-1}p(x_n|\mu)\right]p(x_N|\mu)$$

Now assume the mean is known and we want to infer the precision $\lambda=\frac{1}{\sigma^2}$: **look in the slides**

Mixture of Gaussians

A linear combination of K Gaussians:

$$p(x)=\sum_{k=1}^{K}\pi_k\mathcal{N}(x|\mu_k,\Sigma_k)$$

Where $\sum_{k=1}^{K}\pi_k=1$ and $0\le\pi_k\le1$.

Real data is not Gaussians, so we can think of it as adding many Gaussians with different means and variances. Taking the log:

$$\ln p(X|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\left\{ln\sum_{k=1}^{K}\pi_k\mathcal{N}(x|\mu_k,\Sigma_k)\right\}$$

And this cannot analytically be solved.

Binary variables

A single binary random variable $x \in \{0,1\}$ with $p(x = 1|\mu) = \mu$.

Bernoulli Distribution:

$$Bern(x|\mu) = \mu^x(1-\mu)^{1-x}$$

- $E[x] = \mu$
- $var[x] = \mu(1-\mu)$

Proof:

$E[x] = \sum_{x\in\{0,1\}} xp(x|\mu) = 0(1-\mu) + 1\cdot\mu = \mu$

$var[x] = E[(x - E[x])^2] = E[(x-\mu)^2] = \sum_{x\in\{0,1\}} (x-\mu)^2 p(x|\mu) =$

$\mu^2 p(x=0|\mu) + (1+\mu^2-2\mu)p(x=1|\mu) = \mu(1-\mu)$

Likelihood:

$p(D|\mu) = \prod p(x_n|\mu)$... **look in the slides**

Binomial Distribution

The dist. Of the number m of observations of $x = 1$ given a data set of N observations:

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

The rest is in the slides…

Parametric vs. Non-parametric Density Modeling

Histogram density estimation:

Partition the range of the random variable x into bins of width Δ_i and count the number n_i of observations of x in bin i:

$$p_i = \frac{n_i}{N\Delta_i}$$

Where N is the total number of observations. Usually the bin partition is uniform s.t. $\Delta_i = \Delta$

Nonparametric Density estimators

The probability mass associated with a small region R: $P = \int_R p(x)dx$

With N observations K points within R will be distributed as: $Bin(K,N,P) = \frac{N!}{K!(N-K)!}\ P^K(1-P)^{N-K}$

If R is small enough with volume V then $p(x)$ is constant in $P \cong p(x)V$ so $p(x) = \frac{K}{NV}$

Then we can:

- Fix V and determine K (kernel density estimator)
- Fix K and determine V (knn estimator)

…