CPSC 440/540: Advanced Machine Learning

Gaussians

Winter 2023

End of Part 2 ("Categorical Variables"): Key Concepts

- We discussed categorical density estimation.
  - Model the proportion of times different categories appear.
  - Categorical θ_c parameterization and unnormalized probabilities $\tilde{\theta}_c$.
  - Sampling using the cumulative distribution function (CDF).
- We discussed Monte Carlo for approximating expectations.
  - Generate samples from a model.
  - Compute the average function value on the samples.
- We discussed conjugate priors.
  - For a given likelihood, a prior that leads to posterior in "family" of prior.
  - Conjugate prior for categorical distribution is the Dirichlet distribution.
    - Dirichlet gives a "probability over discrete probabilities".

End of Part 2 (“Categorical Variables”): Key Concepts

- We reviewed standard conditional independence assumptions:
  - Data is IID [given parameters].
  - Data is independent of hyper-parameters given parameters.
  - Discriminative models assume parameters are independent of features.
- We discussed Bayesian learning:
  - Instead of using a single parameter, sum/integrate over all parameters.
  - Prediction using the posterior predictive distribution.
    - And possibly a cost function for Bayesian decision theory.
  - Very-strong protection against overfitting.
- We discussed empirical Bayes:
  - Optimize hyper-parameters using the marginal likelihood.
  - Can optimize a large number of hyper-parameters, without a validation set.
- We discussed hierarchical Bayes:
  - Putting a prior on the prior, which we used to model non-IID grouped data.

End of Part 2 ("Categorical Variables"): Key Concepts

- We discussed multi-class classification.
  - Categorical generalization of sigmoid function is the softmax function.
- We discussed multi-class neural networks.
  - Put softmax on the last layer.
  - Other layers can stay the same, and the same tricks are used/needed.
- We discussed "what have we learned".
  - Layers in CNNs seem to be doing something sensible.
  - But ML models are easily fooled in various ways.
  - And ML models can have harmful biases.

End of Part 2 (“Categorical Variables”): Key Concepts

- We discussed recurrent neural networks (RNNs).
  - Use tied parameters across time to model sequences of different lengths.
    - Makes vanishing/exploding gradient and “forgetting” problems worse.
  - Sequence-to-sequence handles output sequences of unknown lengths.
  - Multi-modal learning considers input and output of different formats.
- We discussed long short term memory (LSTM) models.
  - Include memory cells that are read/written/cleared with gates.
  - Allows modeling longer-range dependencies than standard RNNs.
- We discussed attention.
  - Allows decoder to access information from all encoding steps.
- We discussed transformers.
  - “Fully-connected” attention that forms basis for many modern methods.

Next Topic: Gaussian Density Estimation

Motivating Problem: Cell Phone Battery Life

- Consider modeling battery life between charges:
  - It makes sense to view this as a continuous quantity.
    - Rather than a fixed set of values, the battery life could be any real number.

- Reviews/advertisements will often advertise estimates:

If you want the longest battery life, the iPhone 13 Pro Max is the one to get. In our battery test, the iPhone 13 Pro Max streamed a continuous video at full screen brightness for a whopping **20 hours and 18 minutes**. Nov 11, 2021

https://www.businessinsider.com › ... › Tech › Smartphones

iPhone 13 Pro Max Review: Longest Battery Life and Biggest ...

- We'd like to find the full distribution over charging times.
  - Lets us solve real-world problems like:
    - "If I haven't charged for 18 hours, what is the probability I will make it to 21 hours?"

General Problem: Continuous Density Estimation

- We can view this as density estimation with a continuous variable:
  - Input: n IID samples of continuous values $x^1, x^2, x^3, \ldots, x^n$ from a population.
  - Output: model of probability density for any real number X.
- Continuous density estimation as a picture:

X =

Battery Life
20h18m00s
16h53m42s
21h03m50s
17h33m13s
27h46m28s

$\Rightarrow$ p(x = 20h18m02s) = 0.8

- Watch out: we are estimating the **density** here, not the probability.
  - We could have p(x) > 1.
  - Obtain probabilities by integrating the density over an interval.

Other Applications

- Other applications where continuous density estimation is useful:
  - Modeling sizes (size of food grown in field, birthweight of babies).
  - Modeling times or control values in a manufacturing process.
  - Modeling stock variations or income distributions.
  - Modeling continuous medical measurements (blood pressure).
  - Modeling grades.

- Even with 1 variable there are many possible distributions.
  - More complicated than binary/categorical.

- We'll start with the simple case where we assume data is Gaussian.
  - Also called a "normal" distribution.

Univariate Gaussian

- The Gaussian probability density has the form

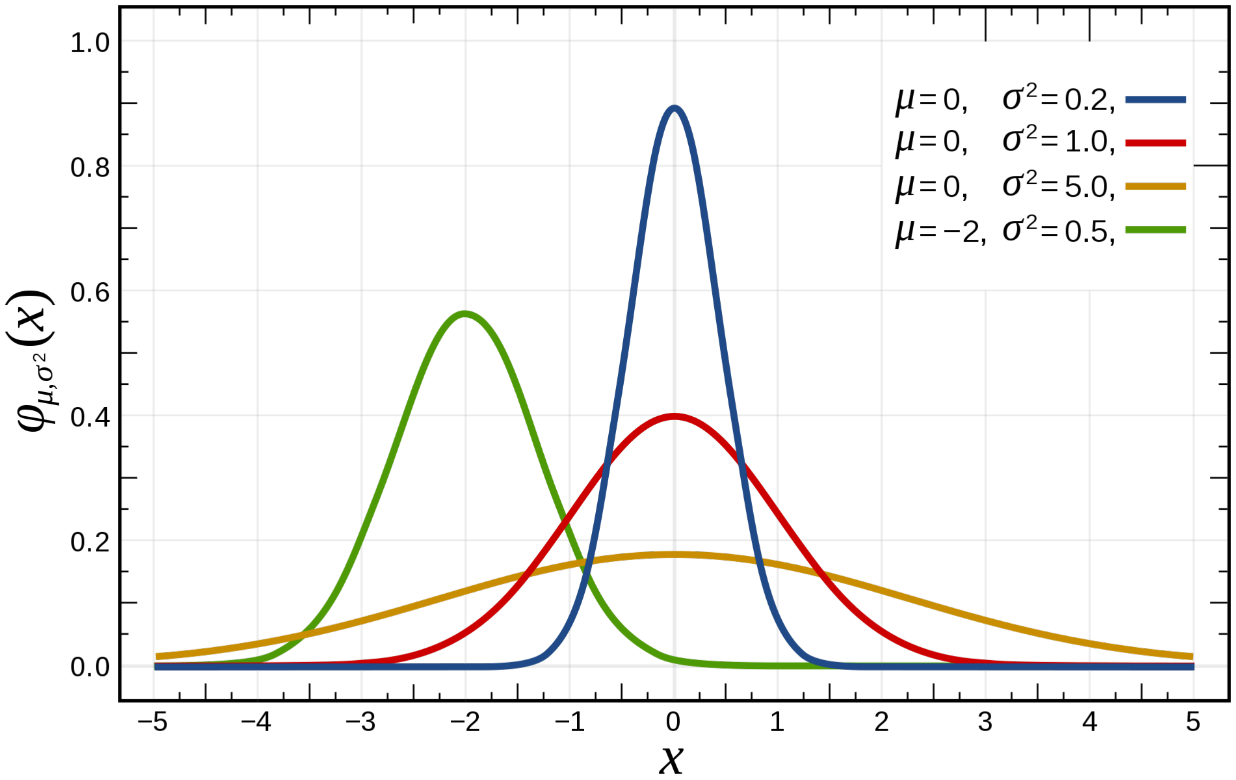
$$p(x^i \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x^i-\mu)^2}{2\sigma^2}\right)$$

  - The mean parameter μ can be any real number.
  - The standard deviation σ can be any positive number.
    - We call σ^2 the variance.
    - Gaussians are also known as normal distributions.
- If we assume x^i follows a Gaussian distribution, we often write:

$$x^i \sim \mathcal{N}(\mu, \sigma^2)$$

"x^i is generated from a normal distribution with mean μ and variance σ^2"

Univariate Gaussian



- Mean parameter μ controls location of center of density.
- Variance parameter σ^2 controls how spread out density is.
  - As $\sigma \to 0$ you get a "spike" at the mean, as $\sigma \to \infty$ you get uniform.

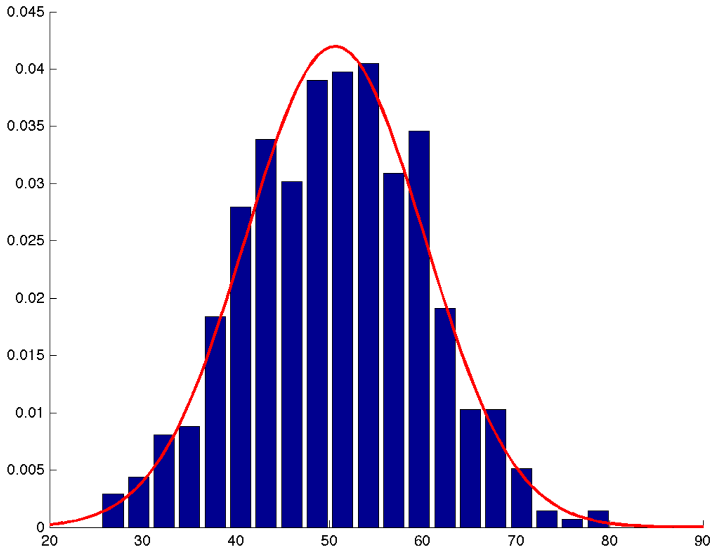
https://en.wikipedia.org/wiki/Normal_distribution

Motivation for Gaussian

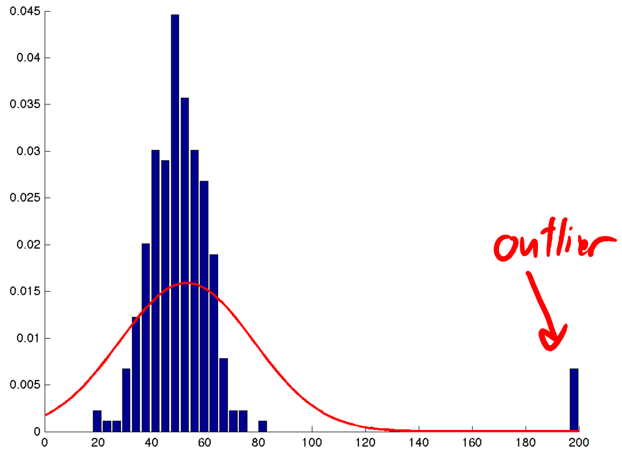
- Why use the Gaussian distribution?
  - Data might actually follow a Gaussian.
    - Good justification if true, but usually false.
  - Central limit theorem: many sums of random variables converge* to Gaussian.
    - Often a bad justification: does not imply data distribution itself converges to a Gaussian.
      - You would have to argue that your data comes from an asymptotic process where CLT applies.
  - The distribution with maximum entropy that fits mean and variance of data.
    - "Makes the least assumptions" while matching the mean and variance of data.
      - We will discuss this later when we discuss the "exponential family".
    - But for complicated problems, just matching means and variances is not enough.
  - Makes many computations and doing theory much easier.
    - The same reason we use a lot of the common distributions.
    - Sometimes Gaussians are "good enough to be useful".
    - Gaussians are common "building blocks" in more advanced methods.

Motivations for not using Gaussians

- Histogram of x^i values with red line being MLE Gaussian density:

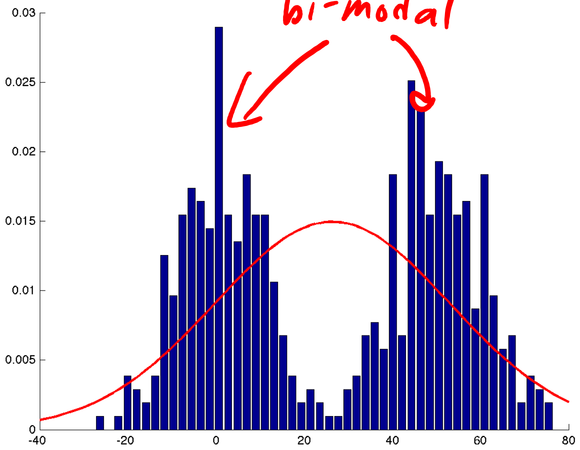


Symmetric around mean, untruncated,
no outliers, uni-modal => ☺

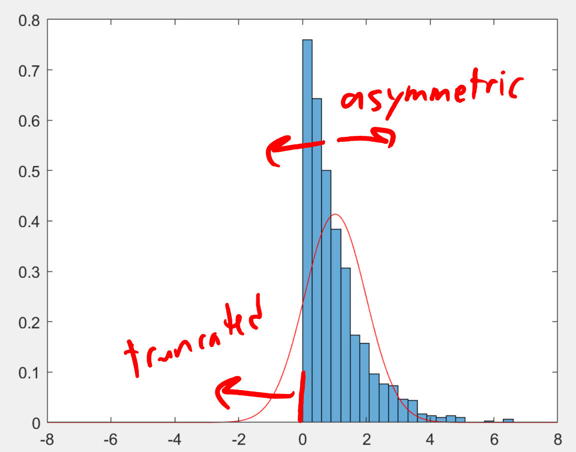


Sensitive to outliers

bi-modal



Cannot model multiple modes



assumes symmetric and not truncated

- Grades usually have all these issues.

Next Topic: Gaussian Inference and Learning

Inference in Univariate Gaussians

- Decoding the mode: find x that maximizes the PDF p(x | μ, σ^2).
  - The mode is the mean μ.
- Computing likelihood of an IID dataset:

$$p(X \mid \mu, \sigma^2) = \prod_{i=1}^{n} p(x^i \mid \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x^i-\mu)^2}{2\sigma^2}\right) = \frac{1}{(\sigma\sqrt{2\pi})^n} \prod_{i=1}^{n} \exp\left(-\frac{(x^i-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{\sum_{i=1}^{n}(x^i-\mu)^2}{2\sigma^2}\right)$$

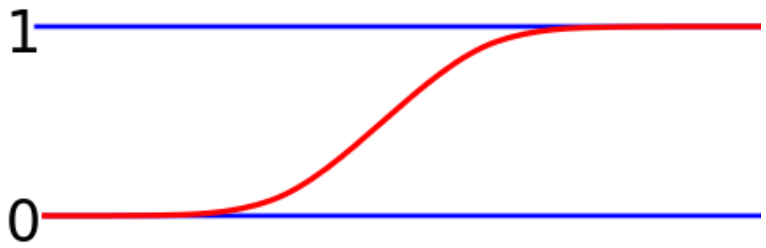
  - Note that the likelihood is a density, not a probability.
- Computing probability that an X lies in an interval:

$$\text{prob}(a \leq x \leq b \mid \mu, \sigma^2) = \int_a^b p(x \mid \mu, \sigma^2)\,dx = \underbrace{\text{prob}(x \leq b \mid \mu, \sigma^2)}_{\text{CDFs}} - \underbrace{\text{prob}(x \leq a \mid \mu, \sigma^2)}_{\text{CDFs}}$$

  - If a=b this is zero: any single x value has probability zero.

Cumulative Distribution Function (CDF)

- We often use F(c) = prob(x ≤ c) = $\int_{-\infty}^{c} p(x)$ to denote the CDF.
  - F(c) is between 0 and 1, giving proportion of times *X* is below *c*.
  - F(c) monotonically increases with 'c'.



- The Gaussian CDF is given by: $F(c) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{c-\mu}{\sigma\sqrt{2}}\right)\right]$
  - The "error function" erf is computed numerically and given by:

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$$

https://en.wikipedia.org/wiki/Cumulative_distribution_function

Sampling with the Inverse CDF ("Quantile") Function

- How can we sample from a continuous density?

- We want to write a function that takes a uniform sample and:
  - 50% of the time it returns a sample in the region where F(c)= 50%.
  - 25% of the time it returns a sample in the region where F(c) = 25%.
  - 75% of the time it returns a sample in the region where F(c) = 75%.
  - 10% of the time it returns a sample in the region where F(c) = 10%.
  - And so on, so the CDF F(c) divides up the interval [0,1].

- The function we want is the inverse of the CDF F^{-1} ("quantile" function):
  - $F^{-1}(u) = c$ for the unique 'c' where $F(c) = u$.
  - Allows sampling from Gaussians and using Monte Carlo with Gaussians.

Inverse Transform Method (Exact 1D Sampling)

- Inverse transform method for exact sampling of a continuous density in 1D:
  1. Sample u uniformly between 0 and 1.
  2. Return $F^{-1}(u)$.

- For Gaussians, we have $F^{-1}(u) = \mu + \sigma\sqrt{2}\text{erf}^{-1}(2u - 1)$.
  - This formula converts uniform u values into samples from a Gaussian.

- Showing that CDF of samples has CDF we want to sample from (for invertible 'F'):

$$
\begin{aligned}
prob(sample \leq c) &= prob(F^{-1}(u) \leq c) && \text{(sample is given by } F^{-1}(u)\text{)} \\
&= prob(F(F^{-1}(u)) \leq F(c)) && \text{(apply strictly-monotonic 'F' to inequality)} \\
&= prob(u \leq F(c)) && \text{(F and } F^{-1} \text{ are inverses)} \\
&= F(c) && \text{(} prob(u \leq y) = y \text{ for uniform 'u')}
\end{aligned}
$$

  - After the inverse transform, we have the CDF of the distribution we want.
- Video on pseudo-randomness and inverse-transform sampling.

MLE for Univariate Gaussian

- We showed that the likelihood for *n* IID examples is given by:

$$p(X \mid \mu, \sigma^2) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{\sum_{i=1}^{n}(x^i - \mu)^2}{2\sigma^2}\right)$$

- To compute the MLE, minimize the NLL (which is convex):

$$-\log p(X \mid \mu, \sigma^2) = n \log \sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x^i - \mu)^2 + \text{constant}$$

- Setting derivative with respect to μ to 0 gives MLE of: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^i$
  - So MLE for the mean is the mean of the samples.
- Plugging in $\hat{\mu}$ and setting derivative with respect to σ to 0 gives: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x^i - \hat{\mu})^2$
  - So MLE for the variance is the variance of the samples.
    - Unless all x^i are equal (then NLL is not bounded below, and MLE does not exist).

Conjugate Prior and Posterior for Mean

- For fixed variance, conjugate prior for mean is Gaussian.

If each $x^i \sim N(\mu, \sigma^2)$ and $\mu \sim N(m, v)$, then $\mu \mid x^1, x^2, \dots, x^n \sim N(\tilde{m}, \tilde{v})$

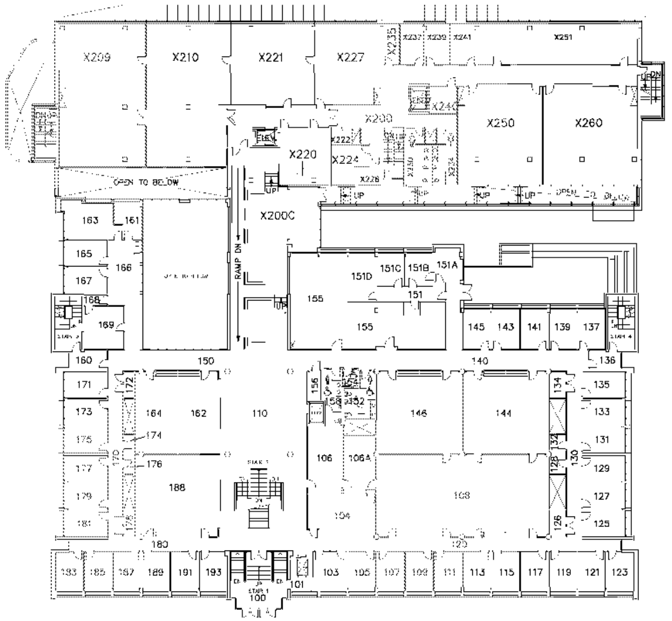
where $\tilde{m} = \frac{vn}{vn+\sigma^2}\hat{\mu}_{MLE} + \frac{\sigma^2}{vn+\sigma^2} m$ and $\tilde{v} = \left(\frac{n}{\sigma^2} + \frac{1}{v}\right)^{-1}$

  - "Self conjugacy" is a **very special** property (a key to usefulness of Gaussians).
    - Derived by using $\propto$ and "completing the square" in exponent (see notes on webpage).
  - Formulas look a bit weird, but consider $\tilde{m}$ and $\tilde{v}$ change as 'n' grows:
    - As *n* grows, posterior mean $\tilde{m}$ converges from prior mean m towards MLE.
    - As *n* grows, posterior variance $\tilde{v}$ converges from prior variance v down to 0.
  - MAP estimate is given by $\tilde{m}$ (it has the highest PDF of the posterior).
  - Posterior predictive is also given by a Gaussian (not obvious, see notes linked on webpage).
    - With mean $\tilde{m}$ and variance $\tilde{v} + \sigma^2$.
    - For complicated Bayeisan inference tasks, can use Monte Carlo by sampling from Gaussian posterior.
- We will come back to MAP/Bayes estimation for variance later.

Next Topic: Multivariate Gaussians

Motivation: Modeling Air Quality

- We want to model "air quality" in different rooms in a building.
- So we measure number of pollutant molecules (PM10, CO, O3, and so on):



Rm 1	Rm 2	Rm 3	Rm 4	Rm 5	Rm 6	Rm 7	Rm 8	Rm 9
0.1	1.4	0.2	1.8	1.0	1.0	0.1	0.1	1.1
0.2	1.3	0.1	1.9	1.1	0.9	0.1	0.1	1.1
0.1	0.3	1.4	2.0	0.7	0.3	0.1	0.2	0.4
0.1	1.1	0.2	2.1	1.1	1.1	0.1	0.3	0.5
2.7	2.6	2.5	5.1	2.4	2.8	3.2	2.5	3.1
0.1	0.4	0.2	1.8	1.3	0.4	0.1	0.4	1.0
0.1	1.2	0.2	1.8	1.4	1.1	0.7	0.7	0.5

- We want to build a model of this data, to identify patterns/problems.
  - Some rooms usually bad air quality, some usually have good air quality.
  - The quality of some rooms may be correlated (rooms are adjacent or share air supply).
  - There are also temporal correlations (we will come back to temporal correlations later).

To Start: Product of Gaussians

- As usual, we could choose to make different dimensions independent

$$X_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

- Then the joint density would be

$$p(x \mid \mu_{1:d}, \sigma_{1:d}) = \prod_{j=1}^{d} p(x_j \mid \mu_j, \sigma_j) \propto \prod_{j=1}^{d} \exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right)$$

$$= \exp\left(-\frac{1}{2}\sum_{j=1}^{d}\frac{1}{\sigma^2}(x_j-\mu_j)^2\right)$$

$$= \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

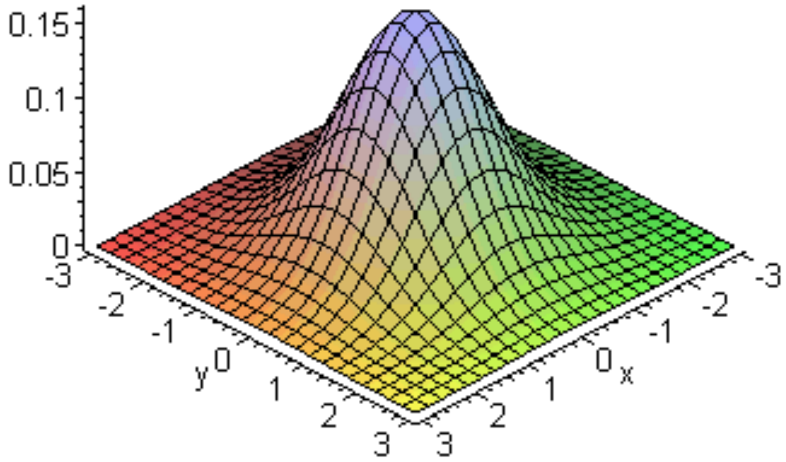
$$\mu = \begin{bmatrix}\mu_1\\ \vdots \\ \mu_d\end{bmatrix}$$

$$\Sigma = \begin{bmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & & \vdots\\ 0 & \cdots & \ddots & \sigma_d^2\end{bmatrix}$$

- General multivariate Gaussian: allow non-diagonal Σ

Multivariate Gaussians

- Many of the nice properties of univariates
  - Closed-form, intuitive MLE / conjugate priors / etc
  - Many nice analytic properties
  - Multivariate central limit theorem
  - …



- Non-diagonal covariance matrix models correlations
  - "Adjacent rooms have similar air qualities"

http://personal.kenyon.edu/hartlaub/MellonProject/Bivariate2.html

Multivariate Gaussian Distribution

If $X \sim \mathcal{N}(\mu, \Sigma)$, $\quad p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\det(\Sigma)|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$

- $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ has $\Sigma \succ 0$, det is the determinant
  - $\Sigma \succ 0$ means that Σ is (strictly) positive definite
    - All eigenvalues are positive
      - Diagonals entries must be positive, but off-diagonal entries can be negative
    - Equivalently, $v^\top \Sigma v > 0$ for all vectors $v \neq 0$
    - Implies there's an A such that $\Sigma = A A^\top$
- Can derive from $X = A Z + \mu$, where $Z_j \sim \mathcal{N}(0,1)$ iid

bonus!

Using change of variables formula $p(x) = \left|\det\left(\frac{\partial z}{\partial x}\right)\right| p(z)$ Jacobian $\left(\frac{\partial Z}{\partial X}\right)_{ij} = \frac{\partial Z_i}{\partial X_j}$

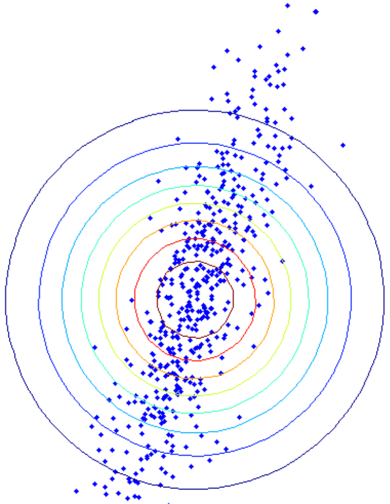
$Z = A^{-1}(X-\mu)$, $\frac{\partial Z}{\partial X} = A^{-1}$; $p(x \mid \mu, A) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\langle A^{-1}(x-\mu), A^{-1}(x-\mu)\rangle\right) |\det A^{-1}|$

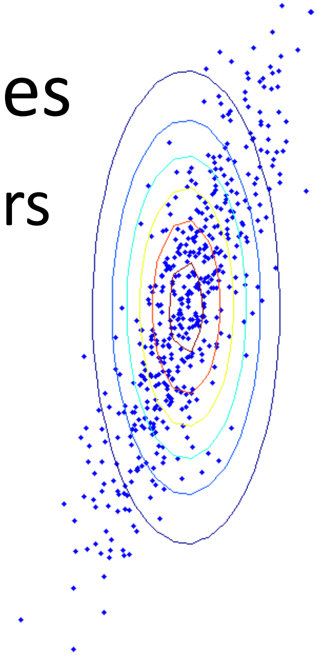
$= \frac{1}{(2\pi)^{d/2} |\det A|} \exp\left(-\frac{1}{2}(x-\mu)^T A^{-T} A^{-1} (x-\mu)\right)$

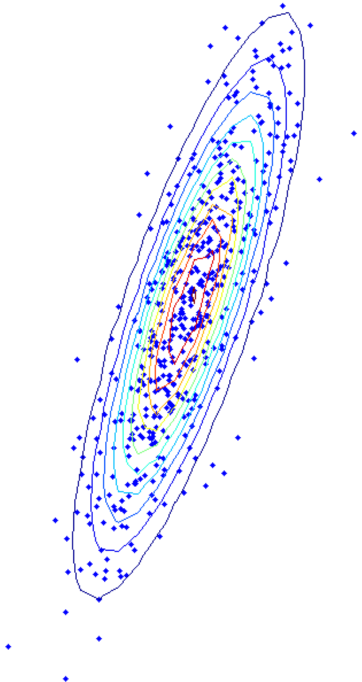
$= \frac{1}{(2\pi)^{d/2} |\det \Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$

Kinds of covariances

- If $\Sigma = \alpha\, I$, level curves of the density are circles
  - Each $X_j \sim \mathcal{N}(0, \alpha)$ is independent; 1 parameter
- If $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is a general diagonal: axis-aligned ellipses
  - Each $X_j \sim \mathcal{N}(0, \sigma_j^2)$ is independent; product of normals; *d* parameters
- If Σ is general, might not be axis-aligned
  - $d(d+1)/2$ parameters
    (not d^2: the matrix is symmetric)





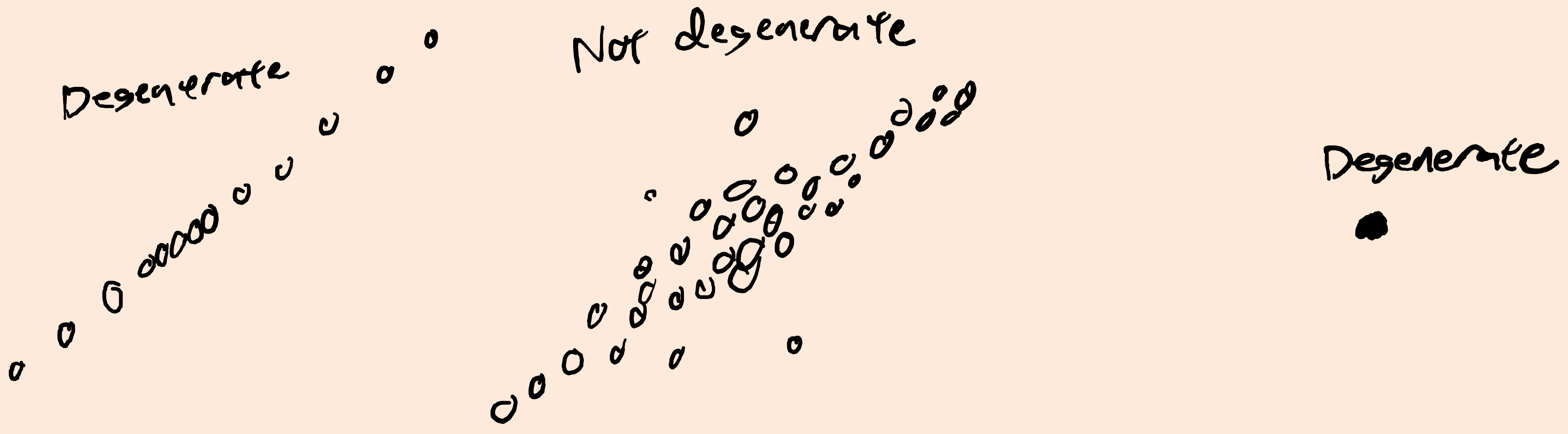


bonus!

Degenerate Gaussians

- If $\det(\Sigma) = 0$ (but still have $\Sigma \succcurlyeq 0$, positive **semi**-definite), we call it a **degenerate** Gaussian
  - Standard density function doesn't exist (divide by 0)

- In 1d, degenerate Gaussians have $\sigma^2 = 0$, a point mass
- In 2d, non-zero probability is along a line (or a point)



Independence structure in Gaussians

- In multivariate Gaussians, $X_j \perp\!\!\!\perp X_{j'}$ iff $\Sigma_{jj'} = 0$
  - If Σ is diagonal, all off-diagonals are 0 and the X_j are all mutually independent

- If $\Sigma_{jj'} \neq 0$, then X_j and $X_{j'}$ are correlated
  - Can be positive or negative

- This means we can model dependencies between all pairs
  - Unlike all the previous "product of […]" distributions we've used
- But no "higher-order" interactions

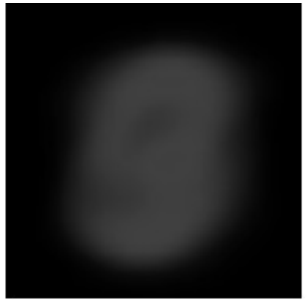
Example: Multivariate Gaussians on MNIST

- Let's try continuous density estimation on handwritten digits

$x^i = \text{vec}\Big($  $\Big)$

Diagonal Σ:

$\hat{\mu} = \text{vec}\Big($  $\Big)$

$\hat{\Sigma} = \text{diag}\Big(\text{vec}\Big($  $\Big)\Big)$

General Σ:

$\hat{\mu}$ is the same (!)

$\hat{\Sigma}$ is big (784 by 784)

Next Topic: Multivariate Gaussian Inference

Inference with Multivariate Gaussians

- How do we use this model?
  - Compute likelihoods with the formula we saw
    - Like 1d Gaussians (and Betas, and any other continuous dist.), likelihood now a **density**
  - Decode the mode: it's again just the mean μ
  - What about marginal distributions, $p(x_j)$?
  - Or conditionals, $p(x_j \mid x_{j'})$?
  - Or sampling from the distribution?

- Gaussians have many nice properties that make computations easy
  - We'll mostly introduce them as we go

Affine Transformations

- If $X \sim \mathcal{N}(\mu, \Sigma)$, then $X + b \sim \mathcal{N}(\mu + b, \Sigma)$

- If $X \sim \mathcal{N}(\mu, \Sigma)$, then $\mathrm{A}\, X + b \sim \mathcal{N}(A\, \mu + b, A\, \Sigma\, \mathrm{A}^\mathsf{T})$
  – $A\, \Sigma\, \mathrm{A}^\mathsf{T}$ might be singular, in which case $\mathrm{A}\, X + b$ is degenerate!
    - e.g. A = 0, or if X is 1d and A is 5×1…

- This gives us a nice sampling algorithm:
  – Sample *d* independent standard normals, $Z_j \sim \mathcal{N}(0, 1)$
  – Return $A\, Z + \mu \sim \mathcal{N}(\mu, AA^\mathsf{T})$
    - Find an A so that $AA^\mathsf{T} = \Sigma$, e.g. Cholesky factorization (`np.linalg.cholesky`)

Marginalizing Gaussians

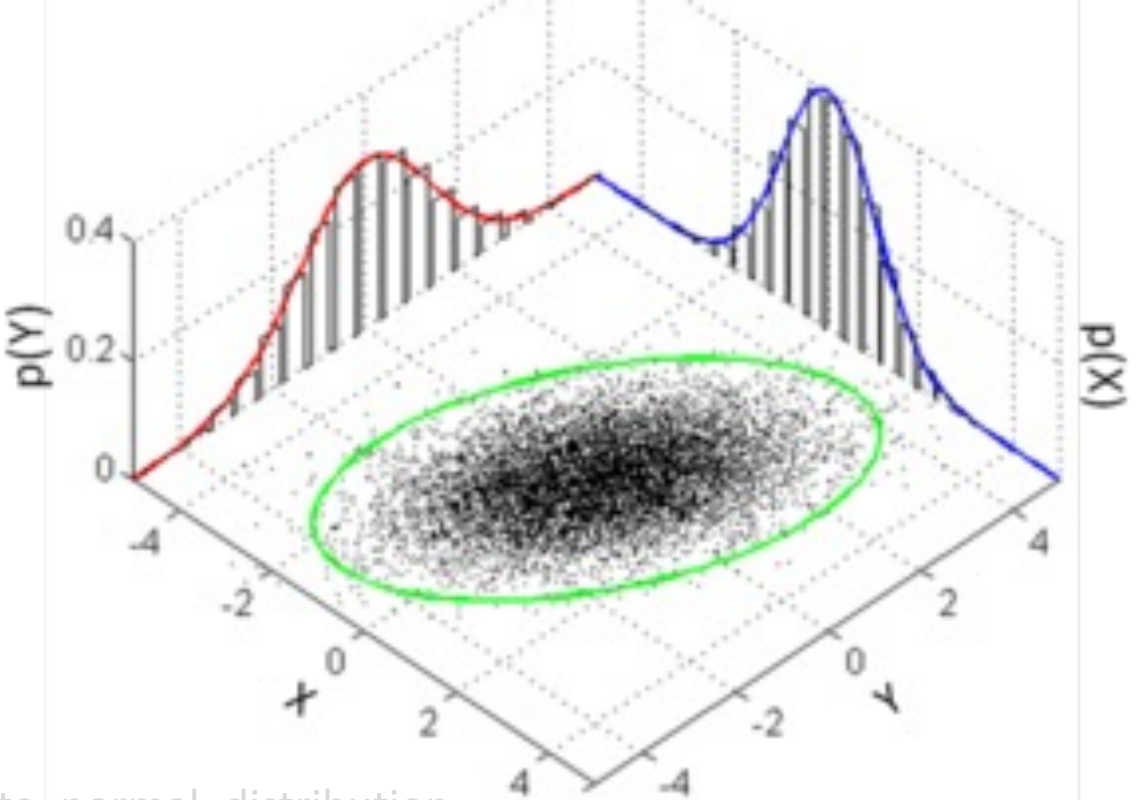
- If we have a joint on $(X_1, X_2, \ldots, X_d)$, might want just X_j
- $p(x_j) = \int \mathrm{d}x_1 \cdots \int \mathrm{d}x_{j-1} \int \mathrm{d}x_{j+1} \cdots \int \mathrm{d}x_d \, p(x \mid \mu, \Sigma)$
  - …but we can skip the integration by thinking a bit!
- Let's partition our variables, $\begin{bmatrix} X \\ Z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Z \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{bmatrix} \right)$
- Now notice that
  $X = \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} X \\ Z \end{bmatrix}$
- and so
  $X \sim \mathcal{N}\left(\begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} \mu_X \\ \mu_Z \end{bmatrix}, \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{bmatrix} \begin{bmatrix} I & 0 \end{bmatrix}^\mathsf{T} \right)$
  $X \sim \mathcal{N}(\mu_X, \Sigma_{XX})$

$$\begin{bmatrix} x_1 \\ x_2 \\ z_1 \\ z_2 \\ z_3 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0.6 \\ -1.2 \\ 1.4 \\ 0.1 \\ 7.3 \end{bmatrix}, \begin{bmatrix} 1.3 & -0.1 & -0.2 & 0.4 & 0 \\ -0.1 & 3.6 & 0.1 & 0.3 & -0.5 \\ -0.2 & 0.1 & 8.1 & -0.2 & 1.4 \\ 0.4 & 0.3 & -0.2 & 1.8 & -0.7 \\ 0 & -0.5 & 1.4 & -0.7 & 2.3 \end{bmatrix} \right)$$

μ_X μ_Z Σ_{XX} Σ_{XZ} $\Sigma_{ZX} = \Sigma_{XZ}^T$ Σ_{ZZ}

Marginalizing Gaussians

- If $\begin{bmatrix} X \\ Z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Z \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{bmatrix}\right)$ then $X \sim \mathcal{N}(\mu_X, \Sigma_{XX})$
- i.e. we can just ignore a subset of the variables



https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Conditioning in Gaussians

- If $\begin{bmatrix} X \\ Z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Z \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{bmatrix}\right)$, what's $X \mid Z$?
- By doing a bunch of linear algebra (see PML1 7.3.5), you get

$$X \mid Z \sim \mathcal{N}(\mu_{X|Z}, \Sigma_{X|Z})$$
$$\mu_{X|Z} = \mu_X + \Sigma_{XZ}\Sigma_{ZZ}^{-1}(Z - \mu_Z)$$
$$\Sigma_{X|Z} = \Sigma_{XX} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$$

- If you know $Z = z$, distribution of X is still (a different) Gaussian
- If $\Sigma_{XZ} = 0$, get $X \mid Z \sim \mathcal{N}(\mu_X, \Sigma_X)$, and so then $X \perp\!\!\!\perp Z$
- Notice that $\Sigma_{X|Z}$ doesn't depend on the particular value of Z!

Summary

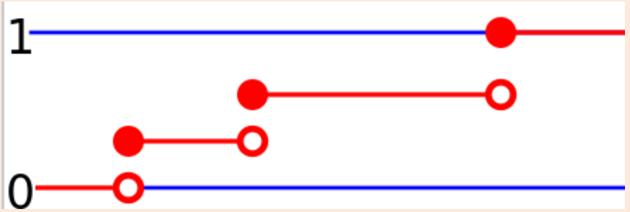
- Gaussian density estimation:
  - Modeling continuous variable samples, assuming it follows a Gaussian.
  - We use Gaussians because they have lots of nice properties.
  - But Gaussians assume symmetric, no outliers, no truncation, uni-modal.
- Mean and variance parameterization of Gaussians:
  - Mean specifies center of distribution.
  - Variance specifies spread of distribution.
- Inverse transform method for sampling:
  - Apply the "inverse" of the CDF to uniform samples to generate samples.
- MLE and MAP for Gaussians:
  - MLE is given by mean and variance of samples.
  - Conjugate prior for mean is another Gaussian.
    - MAP moves between mean of samples and prior mean.
    - Posterior predictive is also Gaussian in this case.
- Multivariate Gaussian for vectors.
  - Mean vector and positive-definite covariance.
  - Diagonal covariance ⇔ product of independent Gaussians.
  - Correlations with off-diagonal entries.
- Inference with multivariate Gaussians
  - Affine transforms of Gaussians are Gaussian.
  - Can use that to sample.
  - Marginals, conditionals are also Gaussians.

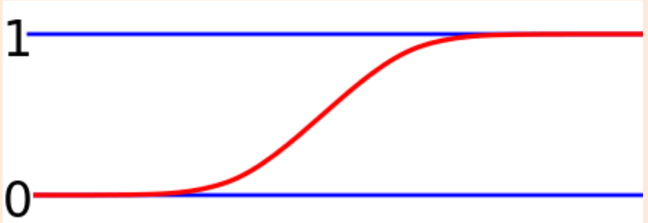
- Next time: learning about how to learn multivariate Gaussians.

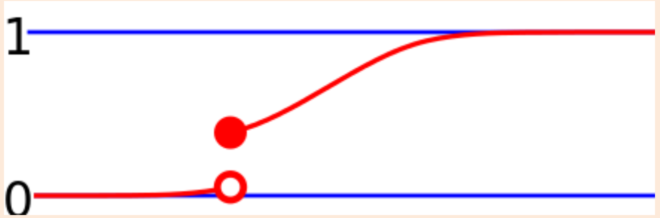
Cumulative Distribution Function (CDF)

bonus!

- CDF can be used for discrete and continuous variables (and mixed).







- We can generalize the quantile function to non-invertible case.

https://en.wikipedia.org/wiki/Cumulative_distribution_function bonus!

Quantile Function – Non-Invertible Case

- If the CDF 'F' is not invertible, we define the quantile F^{-1} as:

$$F^{-1}(u) = \inf\{c \mid F(c) \geq u\}$$

- "Smallest value 'c' such that F(c) is bigger than u."
  - See notes on max and argmax if you have not seen 'inf' before.
    - It's a variant on 'min' that is defined in more cases.

- If 'F' is invertible at this 'c', this gives the usual inverse.
  - But this more-general definition handles non-invertible points.
    - For example, the CDF is not invertible for categorical variables at the "jumps" in CDF.
      - Many values of 'u' are mapped to by the same 'c'.