CPSC 540: Machine Learning

Gaussians

Mark Schmidt

University of British Columbia

Winter 2020

Last Time: Density Estimation

- We started discussing density estimation:

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \quad \tilde{X} = \begin{bmatrix} ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \end{bmatrix}$$

  - What is probability (or PDF) of $[1\ 0\ 1\ 1]$?
    - With model you do inference: test likelihood, sample, conditionals,. . .

- We disucssed product of independent distributions:
  - Just model each column independently (as Bernoulli or categorical).
  - Maybe with Laplace smoothing.
- We discussed general discrete distribution
  - Have one parameter for each of the k^d possible vectors.
  - Not limited in complexity like "product of independent" but leads to overfitting.

Univariate Gaussian

- Consider the case of a continuous variable $x \in \mathbb{R}$:
  - Grades, amounts, velocities, temperatures, and so on.

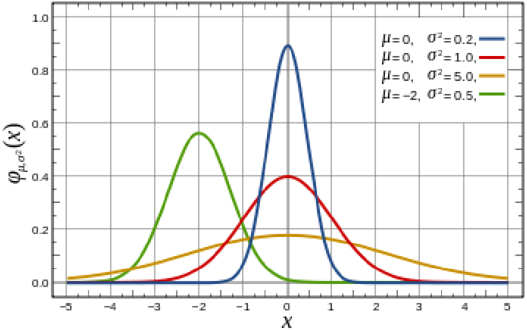
$$X = \begin{bmatrix} 0.53 \\ 1.83 \\ -2.26 \\ 0.86 \end{bmatrix}.$$

- Even with 1 variable there are many possible distributions.
- Most common is the Gaussian (or "normal") distribution:

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

for mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$ (or variance σ^2).

Univariate Gaussian



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

- Mean parameter μ controls location of center of density.
- Variance parameter σ^2 controls how spread out density is.

Univariate Gaussian

- Why use the Gaussian distribution?
  - Data might actually follow Gaussian.
    - Good justification if true, but usually false.
  - Central limit theorem: mean estimators converge in distribution to a Gaussian.
    - Bad justification: doesn't imply data distribution converges to Gaussian.
  - Distribution with maximum entropy that fits mean and variance of data (bonus).
    - "Makes the least assumptions" while matching mean and variance of data.
    - But for complicated problems, just matching mean and variance isn't enough.
  - Closed-form maximum likelihood estimate (MLE).
    - MLE for the mean is the mean of the data ("sample mean" or "empirical mean").
    - MLE for the variance is the variance of the data ("sample variance").
    - A lot of other nice properties that make computation/theory easy.

Univariate Gaussian (MLE for Mean)

- Gaussian likelihood for an example x^i is

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

- So the negative log-likelihood for n IID examples is

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

- Setting derivative with respect to μ to 0 gives MLE of

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^i \quad (\text{for any } \sigma > 0),$$

so the MLE is the mean of the samples.

Univariate Gaussian (MLE for Variance)

- Gaussian likelihood for an example x^i is

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

- So the negative log-likelihood for n IID examples is

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

- Plugging in $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^i$ and setting derivative with respect to σ to zero gives

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x^i - \hat{\mu})^2, \quad \text{(variance of the samples)}$$

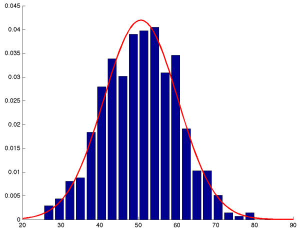
unless all x^i are equal (then NLL is not bounded below and MLE doesn't exist).

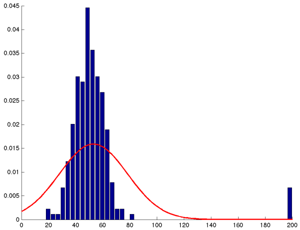
Alternatives to Univariate Gaussian

- Why not the Gaussian distribution?
  - Negative log-likelihood is a quadratic function of μ,

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

  so as with least squares the Gaussian is not robust to outliers.

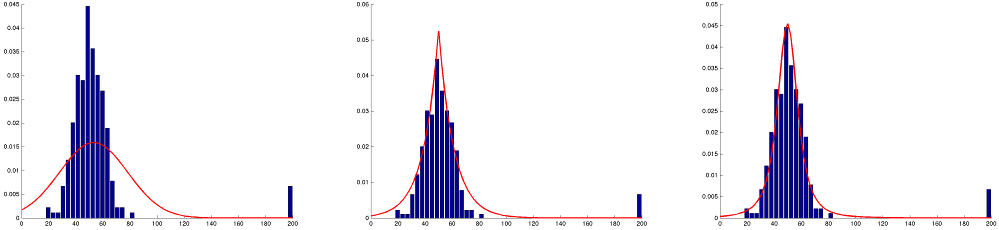




- This is a histogram of the x^i values, and the red line is the estimated density.
- We say Gaussian is "light-tailed": assumes most data is close to mean.

Alternatives to Univariate Gaussian

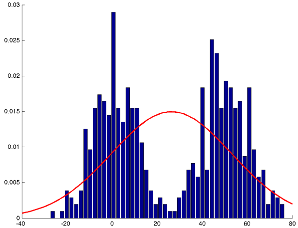
- Robust: Laplace distribution or student's t-distribution



- "Heavy-tailed": has non-trivial probability that data is far from mean.

Alternatives to Univariate Gaussian

- Gaussian distribution is unimodal.



- Laplace and student t are also unimodal so don't fix this issue.
  - Next time we'll discuss "mixture models" that address this.

Outline

Product of Independent Gaussians

- If we have d variables, we could make each follow an independent Gaussian,

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

- In this case the joint density can be written in matrix notation as

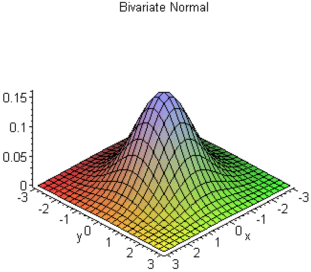
$$\begin{aligned}
\prod_{j=1}^{d} p(x_j^i \mid \mu_j, \sigma_j^2) &\propto \prod_{j=1}^{d} \exp\left(-\frac{(x_j^i - \mu_j)^2}{2\sigma_j^2}\right) \\
&= \exp\left(-\frac{1}{2}\sum_{j=1}^{d} \frac{1}{\sigma_j^2}(x_j^i - \mu_j)^2\right) && (e^a e^b = e^{a+b}) \\
&= \exp\left(-\frac{1}{2}(x^i - \mu)^T \Sigma^{-1}(x - \mu)\right) && \text{(matrix notation)}
\end{aligned}$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_d)$ and Σ is a diagonal matrix with diagonal elements σ_j^2.

- Distributions with this form are a special case of the multivariate Gaussian.

Multivariate Gaussian Distribution

- A $d > 1$ generalization of unvariate Gaussian is the multivariate normal/Gaussian,



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

- This maintains many of the nice properties of univariate Gaussians.
  - Closed-form intuitive MLE, many analytic properties, maximum entropy property.

Multivariate Gaussian Distribution

- The probability density for the multivariate Gaussian is given by

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

  where $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$ and $\Sigma \succ 0$, and $|\Sigma|$ is the determinant.

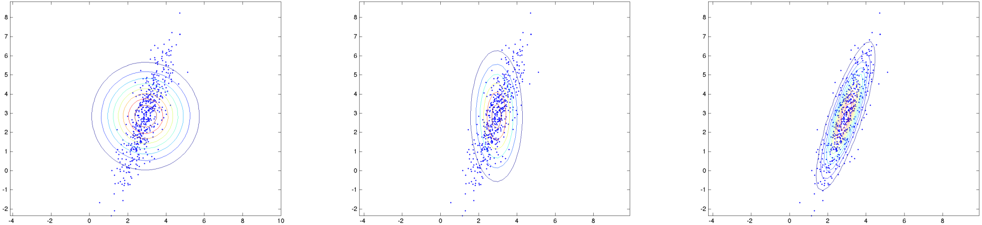
- Where does this wonky formula come from?
  - Consider a product of independent Gaussians, $z_j^i \sim \mathcal{N}(0, 1)$.
  - Then perform a linear transformation, $x^i = Az^i + \mu$.
    - If we define $\Sigma = AA^T$, multivariate Gaussian is PDF of transformed variables.
    - Derivation in bonus slides.

- If $|\Sigma| = 0$ we say the Gaussian is degenerate (bonus).
  - Transformed variables x^i don't span the full space.

Product of Independent Gaussians

- The effect of a diagonal Σ on the multivariate Gaussian:
  - If $\Sigma = \alpha I$ the level curves are circles: 1 parameter.
  - If $\Sigma = D$ (diagonal) then axis-aligned ellipses: d parameters.
    - We saw that this is equivalent to using a product of independent Gaussians.
  - If Σ is dense they do not need to be axis-aligned: $d(d+1)/2$ parameters.

(by symmetry, we only need upper-triangular part of Σ)



- Diagonal Σ assumes features are independent, dense Σ models dependencies.

MLE for Multivariate Gaussian (Mean Vector)

- With a multivariate Gaussian we have

$$p(x^i \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^i - \mu)^\top \Sigma^{-1}(x^i - \mu)\right),$$

so up to a constant our negative log-likelihood for n examples x^i is

$$\frac{1}{2}\sum_{i=1}^{n}(x^i - \mu)^\top \Sigma^{-1}(x^i - \mu) + \frac{n}{2}\log|\Sigma|.$$

- This is a strongly-convex quadratic in μ, setting gradient to zero gives

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x^i,$$

which is the unique solution (strong-convexity is due to $\Sigma \succ 0$).

  - MLE for μ is the average along each dimension, and it doesn't depend on Σ.

MLE for Multivariate Gaussians (Covariance Matrix)

- To get MLE for Σ we re-parameterize in terms of precision matrix $\Theta = \Sigma^{-1}$,

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{n}(x^i-\mu)^\top\Sigma^{-1}(x^i-\mu)+\frac{n}{2}\log|\Sigma|\\=&\frac{1}{2}\sum_{i=1}^{n}(x^i-\mu)^\top\Theta(x^i-\mu)+\frac{n}{2}\log|\Theta^{-1}|\end{aligned}$$

- After some tedious linear algebra (in bonus slides) we obtain that this is equal to

$$\frac{n}{2}\text{Tr}(S\Theta)-\frac{n}{2}\log|\Theta|,$$

where:

- S is the empirical covariance of the data, $S = \frac{1}{n}\sum_{i=1}^{n}(x^i-\mu)(x^i-\mu)^\top$.
- Trace operator $\text{Tr}(A)$ is the sum of the diagonal elements of A.

MLE for Multivariate Gaussians (Covariance Matrix)

- So the NLL in terms of the precision matrix Θ and sample covariance S is

$$f(\Theta) = \frac{n}{2}\text{Tr}(S\Theta) - \frac{n}{2}\log|\Theta|, \text{ with } S = \frac{1}{n}\sum_{i=1}^{n}(x^i - \mu)(x^i - \mu)^\top$$

- Weird-looking but has nice properties:
  - $\text{Tr}(S\Theta)$ is linear function of Θ, with $\nabla_\Theta \text{Tr}(S\Theta) = S$.

    (it's the matrix version of an inner-product $s^\top\theta$)
  - Negative log-determinant is strictly-convex and has $\nabla_\Theta \log|\Theta| = \Theta^{-1}$.

    (generalizes $\nabla \log|x| = 1/x$ for for $x > 0$).
- Using these two properties the gradient matrix has a simple form:

$$\nabla f(\Theta) = \frac{n}{2}S - \frac{n}{2}\Theta^{-1}.$$

MLE for Multivariate Gaussians (Covariance Matrix)

- Gradient matrix of NLL with respect to Θ is

$$\nabla f(\Theta) = \frac{n}{2}S - \frac{n}{2}\Theta^{-1}.$$

- The MLE for a given μ is obtained by setting gradient matrix to zero, giving

$$\Theta = S^{-1} \quad \text{or} \quad \Sigma = S = \frac{1}{n}\sum_{i=1}^{n}(x^i - \mu)(x^i - \mu)^\top.$$

- The constraint $\Sigma \succ 0$ means we need positive-definite sample covariance, $S \succ 0$.
    - If S is not invertible, NLL is unbounded below and no MLE exists.
    - This is like requiring "not all values are the same" in univariate Gaussian.
        - In d-dimensions, you need d linearly-independent x^i values.

- For most distributions, the MLEs are not the sample mean and covariance.

MAP Estimation in Multivariate Gaussian (Covariance Matrix)

- A classic regularizer for Σ is to add a diagonal matrix to S and use
$$\Sigma = S + \lambda I,$$
which satisfies $\Sigma \succ 0$ by construction (eigenvalues at least λ).

- This corresponds to a regularizer that penalizes diagonal of the precision,
$$\begin{aligned} f(\Theta) &= \text{Tr}((S + \lambda I)\Theta) - \log|\Theta| \\ &= \text{Tr}(S\Theta + \lambda\Theta) - \log|\Theta| \\ &= \text{Tr}(S\Theta) - \log|\Theta| + \lambda\text{Tr}(\Theta) \\ &= \text{Tr}(S\Theta) - \log|\Theta| + \lambda\sum_{j=1}^{d}|\Theta_{jj}|. \end{aligned}$$

- So this is L1-regularization of diagonals of inverse covariance.
  - But doesn't set to exactly zero as it must be positive-definite.

Graphical LASSO

- A popular generalization called the graphical LASSO,

$$f(\Theta) = \text{Tr}(S\Theta) - \log|\Theta| + \lambda\|\Theta\|_1.$$

where we are using the element-wise L1-norm.

- Gives sparse off-diagonals in Θ.
  - Can solve very large instances with proximal-Newton and other tricks ("QUIC").
- It's common to draw the non-zeroes in Θ as a graph.
  - Has an interpretation in terms on conditional independence (we'll cover this later).

Summary

- Gaussian distribution is a common distribution with many nice properties.
  - Closed-form MLE.
  - But unimodal and not robust.
- Multivariate Gaussian generalizes univariate Gaussian for multiple variables.
  - Parameterized by mean vector μ and positive-definite covariance Σ.
  - Product of independent Gaussians is equivalent to using a diagonal Σ.
  - Closed-form MLE given by sample mean and covariance.
- Next time: a universal model for continuous densities.

MAP for Univariate Gaussian Mean

- Assume $x^i \sim \mathcal{N}(\mu, \sigma^2)$ and assume $\mu \sim \mathcal{N}(\mu_0, 1)$.
- The MAP estimate of μ under these assumptions can be written as

$$\hat{\mu} = \frac{n}{n + \sigma^2}\bar{x} + \frac{\sigma^2}{n + \sigma^2}\mu_0,$$

  where $\bar{x}$ is the sample mean, $\frac{1}{n}\sum_{i=1}^{n} x^i$ (which is the MLE).
- The MAP estimate is a convex combination of the MLE and prior mean μ_0.
  - Regularizer moves us in a straight line away from MLE towards μ_0.

Maximum Entropy and Gaussian

- Consider trying to find the PDF $p(x)$ that
  1. Agrees with the sample mean and sample covariance of the data.
  2. Maximizes entropy subject to these constraints,

$$\max_p \left\{ -\int_{-\infty}^{\infty} p(x) \log p(x) dx \right\}, \quad \text{subject to } \mathbb{E}[x] = \mu, \ \mathbb{E}[(x-\mu)^2] = \sigma^2.$$

- Solution is the Gaussian with mean μ and variance σ^2.
  - Beyond fitting mean/variance, Gaussian makes fewest assumptions about the data.

- This is proved using the convex conjugate.
  - Convex conjugate of Gaussian negative log-likelihood is entropy.
  - Same result holds in higher dimensions for multivariate Gaussian.

Multivariate Gaussian from Univariate Gaussians

- Consider a joint distribution that is the product univariate standard normals:

$$
\begin{aligned}
p(z^i) &= \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z_j^i)^2\right) \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(\frac{1}{2}\langle z^i, z^i\rangle\right).
\end{aligned}
$$

- Now define $x^i = Az^i + \mu$ for some (non-singular) matrix A and vector μ.
- The change of variables formula for multivariate probabilities is

$$
p(x^i) = p(z^i)\left|\frac{\partial z^i}{\partial x^i}\right|.
$$

- Plug in $z^i = A^{-1}(x^i - \mu)$ and $\frac{\partial z^i}{\partial x^i} = A^{-1}$...

Multivariate Gaussian from Univariate Gaussians

- This gives

$$p(x^i \mid \mu, A) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(\frac{1}{2}\langle A^{-1}(x^i - \mu), A^{-1}(x^i\mu)\rangle\right) |\det(A^{-1})|$$
$$= \frac{1}{(2\pi)^{\frac{d}{2}} |\det(A)|} \exp\left(\frac{1}{2}(x^i - \mu) A^{-\top} A^{-1} (x^i - \mu)\right).$$

- Define $\Sigma = AA^\top$ (so $\Sigma^{-1} = A^{-\top}A^{-1}$ and $\det\Sigma = (\det A)^2$) to get

$$p(x^i \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^i - \mu)^\top \Sigma^{-1} (x^i - \mu)\right)$$

- So multivariate Gaussian is an affine transformtation of independent Gaussians.

Degenerate Gaussians

- If $|\Sigma| = 0$, we say the Gaussian is degenerate.
- In this case the PDF only integrates to 1 along a subspace of the original space.
- With $d = 2$ degenerate Gaussians only have non-zero probability along a line (or just one point).

MLE for Multivariate Gaussians (Covariance Matrix)

- To get MLE for Σ we re-parameterize in terms of precision matrix $\Theta = \Sigma^{-1}$,

$$
\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{n}(x^i-\mu)^\top\Sigma^{-1}(x^i-\mu)+\frac{n}{2}\log|\Sigma| \\
&=\frac{1}{2}\sum_{i=1}^{n}(x^i-\mu)^\top\Theta(x^i-\mu)+\frac{n}{2}\log|\Theta^{-1}| && \text{(ok because } \Sigma \text{ is invertible)} \\
&=\frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}\left((x^i-\mu)^\top\Theta(x^i-\mu)\right)+\frac{n}{2}\log|\Theta|^{-1} && \text{(scalar } y^\top Ay = \mathrm{Tr}(y^\top Ay)\text{)} \\
&=\frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}((x^i-\mu)(x^i-\mu)^\top\Theta)-\frac{n}{2}\log|\Theta| && (\mathrm{Tr}(ABC)=\mathrm{Tr}(CAB))
\end{aligned}
$$

- Where the trace $\mathrm{Tr}(A)$ is the sum of the diagonal elements of A.
  - That $\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB)$ when dimensions match is the cyclic property of trace.

MLE for Multivariate Gaussians (Covariance Matrix)

- From the last slide we have in terms of precision matrix Θ that

$$=\frac{1}{2}\sum_{i=1}^{n}\mathsf{Tr}((x^i-\mu)(x^i-\mu)^\top\Theta)-\frac{n}{2}\log|\Theta|$$

- We can exchange the sum and trace (trace is a linear operator) to get,

$$=\frac{1}{2}\mathsf{Tr}\left(\sum_{i=1}^{n}(x^i-\mu)(x^i-\mu)^\top\Theta\right)-\frac{n}{2}\log|\Theta| \qquad \sum_i \mathsf{Tr}(A_iB)=\mathsf{Tr}\left(\sum_i A_iB\right)$$

$$=\frac{n}{2}\mathsf{Tr}\left(\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x^i-\mu)(x^i-\mu)^\top}_{\text{sample covariance '}S\text{'}}\right)\Theta\right)-\frac{n}{2}\log|\Theta|. \qquad \left(\sum_i A_iB\right)=\left(\sum_i A_i\right)B$$

Positive-Definiteness of Θ and Checking Positive-Definiteness

- If we define centered vectors $\tilde{x}^i = x^i - \mu$ then empirical covariance is

$$S = \frac{1}{n}\sum_{i=1}^n (x^i - \mu)(x^i - \mu)^\top = \frac{1}{n}\sum_{i=1}^n \tilde{x}^i(\tilde{x}^i)^\top = \frac{1}{n}\tilde{X}^\top\tilde{X} \succeq 0,$$

  so S is positive semi-definite but not positive-definite by construction.
- If data has noise, it will be positive-definite with n large enough.
- For $\Theta \succ 0$, note that for an upper-triangular T we have

$$\log|T| = \log(\text{prod}(\text{eig}(T))) = \log(\text{prod}(\text{diag}(T))) = \text{Tr}(\log(\text{diag}(T))),$$

  where we've used Matlab notation.
- So to compute $\log|\Theta|$ for $\Theta \succ 0$, use Cholesky to turn into upper-triangular.
  - Bonus: Cholesky fails if $\Theta \succ 0$ is not true, so it checks positive-definite constraint.