Advanced Machine Learning
Lecture 6

Probability Distributions

16.11.2015

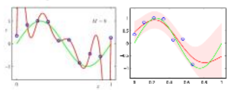
Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de/

leibe@vision.rwth-aachen.de

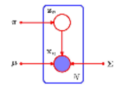
This Lecture: *Advanced Machine Learning*

$f : \mathcal{X} \to \mathbb{R}$

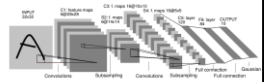
- Regression Approaches
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Gaussian Processes



- Learning with Latent Variables
  - Probability Distributions & Mixture Models
  - Approximate Inference
  - EM and Generalizations



- Deep Learning
  - Neural Networks
  - CNNs, RNNs, RBMs, etc.



Recap: GPs with Noise-free Observations

- Assume our observations are noise-free:

$$\{(\mathbf{x}_n, f_n) \mid n = 1, \ldots, N\}$$

  - Joint distribution of the training outputs f and test outputs $\mathbf{f}_\star$ according to the prior:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_\star \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) & K(X,X_\star) \\ K(X_\star,X) & K(X_\star,X_\star) \end{bmatrix}\right)$$

  - Calculation of posterior corresponds to conditioning the joint Gaussian prior distribution on the observations:

$$\mathbf{f}_\star | X_\star, X, \mathbf{f} \sim \mathcal{N}(\bar{\mathbf{f}}_\star, \operatorname{cov}[\mathbf{f}_\star]) \qquad \bar{\mathbf{f}}_\star = \mathbb{E}[\mathbf{f}_\star | X, X_\star, \mathbf{f}]$$

  - with:

$$\begin{aligned} \bar{\mathbf{f}}_\star &= K(X_\star, X) K(X,X)^{-1} \mathbf{f} \\ \operatorname{cov}[\mathbf{f}_\star] &= K(X_\star, X_\star) - K(X_\star, X) K(X,X)^{-1} K(X, X_\star) \end{aligned}$$

Recap: GPs with Noisy Observations

- Joint distribution of the observed values and the test locations under the prior:

$$\begin{bmatrix} \mathbf{t} \\ \mathbf{f}_\star \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_\star) \\ K(X_\star,X) & K(X_\star,X_\star) \end{bmatrix}\right)$$

  - Calculation of posterior corresponds to conditioning the joint Gaussian prior distribution on the observations:

$$\mathbf{f}_\star | X_\star, X, \mathbf{t} \sim \mathcal{N}(\bar{\mathbf{f}}_\star, \operatorname{cov}[\mathbf{f}_\star]) \qquad \bar{\mathbf{f}}_\star = \mathbb{E}[\mathbf{f}_\star | X, X_\star, \mathbf{t}]$$

  - with:

$$\begin{aligned} \bar{\mathbf{f}}_\star &= K(X_\star, X) \left(K(X,X) + \sigma_n^2 I\right)^{-1} \mathbf{t} \\ \operatorname{cov}[\mathbf{f}_\star] &= K(X_\star, X_\star) - K(X_\star, X) \left(K(X,X) + \sigma_n^2 I\right)^{-1} K(X, X_\star) \end{aligned}$$

⇒ This is the key result that defines Gaussian process regression!
  - Predictive distribution is Gaussian whose mean and variance depend on test points $X_\star$ and on the kernel $k(\mathbf{x},\mathbf{x}')$, evaluated on X.

Recap: Bayesian Model Selection for GPs

- Goal
  - Determine/learn different parameters of Gaussian Processes
- Hierarchy of parameters
  - Lowest level
    - $\mathbf{w}$ - e.g. parameters of a linear model.
  - Mid-level (hyperparameters)
    - θ - e.g. controlling prior distribution of $\mathbf{w}$.
  - Top level
    - Typically discrete set of model structures $\mathcal{H}_i$.
- Approach
  - Inference takes place one level at a time.

Recap: Model Selection at Lowest Level

- Posterior of the parameters $\mathbf{w}$ is given by Bayes' rule

$$\begin{aligned} p(\mathbf{w}|\mathbf{t}, X, \theta, \mathcal{H}_i) &= \frac{p(\mathbf{t}|X, \mathbf{w}, \theta, \mathcal{H}_i) p(\mathbf{w}|\theta, X, \mathcal{H}_i)}{p(\mathbf{t}|X, \theta, \mathcal{H}_i)} \\ &= \frac{p(\mathbf{t}|X, \mathbf{w}, \mathcal{H}_i) p(\mathbf{w}|\theta, \mathcal{H}_i)}{p(\mathbf{t}|X, \theta, \mathcal{H}_i)} \end{aligned}$$

- with
  - $p(\mathbf{t} \mid X, \mathbf{w}, \mathcal{H}_i)$ likelihood and
  - $p(\mathbf{w} \mid \theta, \mathcal{H}_i)$ prior parameters $\mathbf{w}$,
  - Denominator (normalizing constant) is independent of the parameters and is called marginal likelihood.

$$p(\mathbf{t}|X, \theta, \mathcal{H}_i) = \int p(\mathbf{t}|X, \mathbf{w}, \mathcal{H}_i) p(\mathbf{w}|\theta, \mathcal{H}_i) d\mathbf{w}$$

Recap: Model Selection at Mid Level

- Posterior of parameters θ is again given by Bayes' rule

$$p(\theta|\mathbf{t},X,\mathcal{H}_i) = \frac{p(\mathbf{t}|X,\theta,\mathcal{H}_i)p(\theta|X,\mathcal{H}_i)}{p(\mathbf{t}|X,\mathcal{H}_i)} = \frac{p(\mathbf{t}|X,\theta,\mathcal{H}_i)p(\theta|\mathcal{H}_i)}{p(\mathbf{t}|X,\mathcal{H}_i)}$$

- where
  - The marginal likelihood of the previous level $p(\mathbf{t}|X,\theta,\mathcal{H}_i)$ plays the role of the likelihood of this level.
  - $p(\theta|\mathcal{H}_i)$ is the hyperprior (prior of the hyperparameters)
  - Denominator (normalizing constant) is given by:

$$p(\mathbf{t}|X,\mathcal{H}_i) = \int p(\mathbf{t}|X,\theta,\mathcal{H}_i)p(\theta|\mathcal{H}_i)d\theta$$

Recap: Model Selection at Top Level

- At the top level, we calculate the posterior of the model

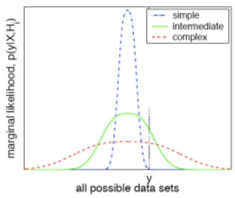
$$p(\mathcal{H}_i|\mathbf{t},X) = \frac{p(\mathbf{t}|X,\mathcal{H}_i)p(\mathcal{H}_i)}{p(\mathbf{t}|X)}$$

- where
  - Again, the denominator of the previous level $p(\mathbf{t}|X,\mathcal{H}_i)$ plays the role of the likelihood.
  - $p(\mathcal{H}_i)$ is the prior of the model structure.
  - Denominator (normalizing constant) is given by:

$$p(\mathbf{t}|X) = \sum_i p(\mathbf{t}|X,\mathcal{H}_i)p(\mathcal{H}_i)$$

Recap: Bayesian Model Selection

- Discussion
  - Marginal likelihood is main difference to non-Bayesian methods
  - It automatically incorporates a trade-off between the model fit and the model complexity:
    - A simple model can only account for a limited range of possible sets of target values - if a simple model fits well, it obtains a high posterior.
    - A complex model can account for a large range of possible sets of target values - therefore, it can never attain a very high posterior.



Topics of This Lecture

- Probability Distributions
  - Bayesian Estimation Reloaded
- Binary Variables
  - Bernoulli distribution
  - Binomial distribution
  - Beta distribution
- Multinomial Variables
  - Multinomial distribution
  - Dirichlet distribution
- Continuous Variables
  - Gaussian distribution
  - Gamma distribution
  - Student's t distribution
  - Exponential Family

Motivation

- Recall: Bayesian estimation

$$p(x|X) = \int p(x|\theta)\frac{p(X|\theta)p(\theta)}{\int p(X|\theta')p(\theta')\mathrm{d}\theta'}\mathrm{d}\theta$$

  - So far, we have only done this for Gaussian distributions, where the integrals could be solved analytically.
  - Now, let's also examine other distributions…

Teaser: Conjugate Priors

- Problem: How to evaluate the integrals?
  - We will see that if likelihood and prior have the same functional form $c\cdot f(x)$, then the analysis will be greatly simplified and the integrals will be solvable in closed form.

$$p(X|\theta)p(\theta) = \prod_{x_n} c_1 f(x_n,\theta)c_2 f(\theta,\alpha) = \prod_{x_n} c f(x_n,\theta,\alpha)$$

  - Such an algebraically convenient choice is called a conjugate prior. Whenever possible, we should use it.
  - To do this, we need to know for each probability distribution what is its conjugate prior. ⇒ *Topic of this lecture.*
- What to do when we cannot use the conjugate prior?
  - ⇒ *Use approximate inference methods. Next lecture…*

Topics of This Lecture

- Probability Distributions
  - Bayesian Estimation Reloaded
- **Binary Variables**
  - Bernoulli distribution
  - Binomial distribution
  - Beta distribution
- Multinomial Variables
  - Multinomial distribution
  - Dirichlet distribution
- Continuous Variables
  - Gaussian distribution
  - Gamma distribution
  - Student's t distribution
  - Exponential Family

Binary Variables

- **Example: Flipping a coin**
  - Binary random variable $x \in \{0,1\}$
  - Outcome heads: $x = 1$
  - Outcome tails: $x = 0$
  - Denote probability of landing heads by parameter μ

$$p(x=1|\mu) = \mu$$

- **Bernoulli distribution**
  - Probability distribution over x:

$$\begin{aligned}\mathrm{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu)\end{aligned}$$

The Binomial Distribution

- **Now consider N coin flips**
  - Probability of landing m heads: $p(m \text{ heads}|N,\mu)$
- **Binomial distribution**

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

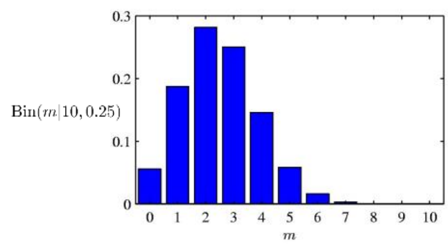
  - Properties

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\mathrm{Bin}(m|N,\mu) = N\mu$$

$$\mathrm{var}[m] \equiv \sum_{m=0}^{N}(m-\mathbb{E}[m])^2\,\mathrm{Bin}(m|N,\mu) = N\mu(1-\mu)$$

  - Note: Bernoulli is a special case of the Binomial for $n = 1$.

Binomial Distribution: Visualization



Parameter Estimation: Maximum Likelihood

- **Maximum Likelihood for Bernoulli**
  - Given a data set $\mathcal{D} = \{x_1,\ldots,x_N\}$ of observed values for x.
  - Likelihood

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N}\mu^{x_n}(1-\mu)^{1-x_n}$$

$$\log p(\mathcal{D}|\mu) = \sum_{n=1}^{N}\log p(x_n|\mu) = \sum_{n=1}^{N}\{x_n\log\mu + (1-x_n)\log(1-\mu)\}$$

- **Observation**
  - The log-likelihood depends on the observations x_n only through their sum.

  ⇒ $\Sigma_n x_n$ is a sufficient statistic for the Bernoulli distribution.

ML for Bernoulli Distribution

$$\log p(\mathcal{D}|\mu) = \sum_{n=1}^{N}\{x_n\log\mu + (1-x_n)\log(1-\mu)\}$$

$$\nabla_\mu \log p(\mathcal{D}|\mu) = \frac{1}{\mu}\sum_{n=1}^{N}x_n - \frac{1}{1-\mu}\sum_{n=1}^{N}(1-x_n) \stackrel{!}{=} 0$$

$$(1-\mu)\sum_{n=1}^{N}x_n = \mu\sum_{n=1}^{N}(1-x_n)$$

$$\sum_{n=1}^{N}x_n - \cancel{\mu\sum_{n=1}^{N}x_n} = N\mu - \cancel{\mu\sum_{n=1}^{N}x_n}$$

- **ML estimate:**

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n$$

ML for Bernoulli Distribution

- Maximum Likelihood estimate

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N} \qquad \text{for } m \text{ heads } (x_n = 1)$$

- Discussion
  - Consider a data set $\mathcal{D} = \{1,1,1\}$. $\quad \rightarrow \mu_{\mathrm{ML}} = \frac{3}{3} = 1$

  ⇒ Prediction: *all* future tosses will land head up!

  ⇒ Overfitting to $\mathcal{D}$!

Bayesian Bernoulli: First Try

- Bayesian estimation
  - We can improve the ML estimate by incorporating a prior for μ.
  - How should such a prior look like?
  - Consider the Bernoulli/Binomial form

$$p(\mathcal{D}|\mu) \propto \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

  - If we choose a prior with the same functional form, then we will get a closed-form expression for the posterior; otherwise, a difficult numerical integration may be necessary.
  - Most general form here:

$$p(\mu|a,b) \propto \mu^{a}(1-\mu)^{b}$$

  - This algebraically convenient choice is called a conjugate prior.

The Beta Distribution

- Beta distribution
  - Distribution over $\mu \in [0,1]$:

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

  - Where $\Gamma(x)$ is the gamma function

$$\Gamma(x) \equiv \int_0^{\infty} u^{x-1}e^{-u}\mathrm{d}u$$

  for which $\Gamma(x+1) = x!$ iff x is an integer.

  ⇒ $\Gamma(x)$ is a continuous generalization of the factorial.
  - The Beta distribution generalizes the Binomial to arbitrary values of a and b, while keeping the same functional form.
  - It is therefore a conjugate prior for the Bernoulli and Binomial.

Beta Distribution

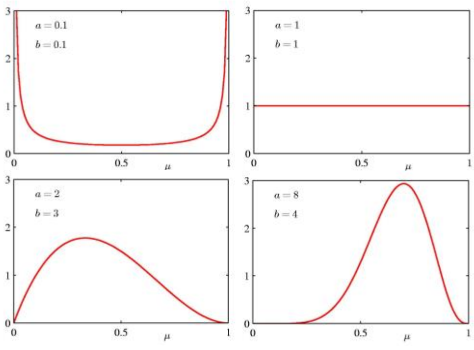
- Properties
  - In general, the Beta distribution is a suitable model for the random behavior of percentages and proportions.
  - Mean and variance

$$\mathbb{E}[\mu] = \frac{a}{a+b}$$

$$\mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

  - The parameters a and b are often called hyperparameters, because they control the distribution of the parameter μ.
  - General observation: if a distribution has K parameters, then the conjugate prior typically has K+1 hyperparameters.

Beta Distribution: Visualization



Bayesian Bernoulli

- Bayesian estimate

$$\begin{aligned} p(\mu|a_0,b_0,\mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0,b_0) \\ &= \left(\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}\right)\mathrm{Beta}(\mu|a_0,b_0) \\ &\propto \mu^{m+a_0-1}(1-\mu)^{(N-m)+b_0-1} \\ &\propto \mathrm{Beta}(\mu|a_N,b_N) \end{aligned}$$

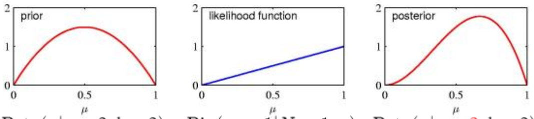
  - This is again a Beta distribution with the parameters

$$a_N = a_0 + m \qquad b_N = b_0 + (N-m)$$

  ⇒ We can interpret the hyperparameters a and b as an effective number of observations for $x = 1$ and $x = 0$, respectively.
  - Note: a and b need not be integers!

Sequential Estimation

- Prior · Likelihood = Posterior
  - The posterior can act as a prior if we observe additional data.
  - The number of effective observations increases accordingly.
- Example: Taking observations one at a time



$\mathrm{Beta}(\mu|a=2, b=2)$ $\mathrm{Bin}(m=1|N=1,\mu)$ $\mathrm{Beta}(\mu|a=3, b=2)$

$\Rightarrow$ This sequential approach to learning naturally arises when we take a Bayesian viewpoint.

Properties of the Posterior

- Behavior in the limit of infinite data
  - As the size of the data set, N, increases

$$\begin{aligned} a_N &= a_0 + m \to m \\ b_N &= b_0 + N - m \to N - m \\ \mathbb{E}[\mu] &= \frac{a_N}{a_N + b_N} \to \frac{m}{N} = \mu_{\mathrm{ML}} \\ \mathrm{var}[\mu] &= \frac{a_N b_N}{(a_N+b_N)^2(a_N+b_N+1)} \to 0 \end{aligned}$$

$\Rightarrow$ As expected, the Bayesian result reduces to the ML result.

Prediction under the Posterior

- Predict the outcome of the next trial
  - "What is the probability that the next coin toss will land heads up?"

$\Rightarrow$ Evaluate the predictive distribution of x given the observed data set $\mathcal{D}$:

$$\begin{aligned} p(x=1|a_0,b_0,\mathcal{D}) &= \int_0^1 p(x=1|\mu)p(\mu|a_0,b_0,\mathcal{D})\,\mathrm{d}\mu \\ &= \int_0^1 \mu p(\mu|a_0,b_0,\mathcal{D})\,\mathrm{d}\mu \\ &= \mathbb{E}[\mu|a_0,b_0,\mathcal{D}] = \frac{a_N}{a_N+b_N} \end{aligned}$$

  - Simple interpretation: total fraction of observations that correspond to $x = 1$.

Topics of This Lecture

Multinomial Variables

- Multinomial variables
  - Variables that can take one of K possible distinct states
  - Convenient: 1-of-K coding scheme: $\mathbf{x} = (0,0,1,0,0,0)^{\mathrm{T}}$
- Generalization of the Bernoulli distribution
  - Distribution of $\mathbf{x}$ with K outcomes

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

with the constraints

$$\forall k: \mu_k \geqslant 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$

Multinomial Variables

- Properties
  - Distribution is normalized

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1$$

  - Expectation

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}} = \boldsymbol{\mu}$$

  - Likelihood given a data set $\mathcal{D} = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$:

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}$$

where m_k is the number of cases for which $\mathbf{x}_n$ has output k.

ML Parameter Estimation

- Maximum Likelihood solution for μ
  - Need to maximize

$$\log p(\mathcal{D}|\mu) = \log \prod_{k=1}^{K} \mu_k^{m_k} = \sum_{k=1}^{K} m_k \log \mu_k$$

Under the constraint $\sum_k \mu_k = 1$

- Solution with Lagrange multiplier

$$\arg\max_{\mu} \sum_{k=1}^{K} m_k \log \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

  - Setting the derivative to zero yields

$$\mu_k = -m_k / \lambda \qquad \mu_k^{\mathrm{ML}} = \frac{m_k}{N}$$

The Multinomial Distribution

- Multinomial Distribution
  - Joint distribution over $m_1,\dots,m_K$ conditioned on μ and N

$$\mathrm{Mult}(m_1, m_2, \dots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

with the normalization coefficient

$$\binom{N}{m_1 m_2 \dots m_K} = \frac{N!}{m_1! m_2! \dots m_K!}$$

  - Properties

$$\begin{aligned} \mathbb{E}[m_k] &= N\mu_k \\ \mathrm{var}[m_k] &= N\mu_k(1-\mu_k) \\ \mathrm{cov}[m_j m_k] &= -N\mu_j\mu_k \end{aligned}$$

Bayesian Multinomial

- Conjugate prior for the Multinomial
  - Introduce a family of prior distributions for the parameters $\{\mu_k\}$ of the Multinomial.
  - The conjugate prior is given by

$$p(\mu|\alpha) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

with the constraints

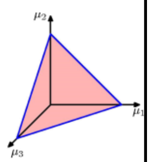
$$\forall k : 0 \leq \mu_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$

The Dirichlet Distribution

- Dirichlet Distribution
  - Multivariate generalization of the Beta distribution

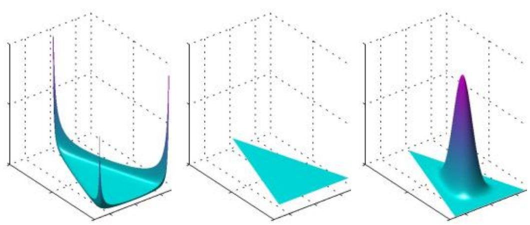
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \quad \text{with} \quad \alpha_0 = \sum_{k=1}^{K} \alpha_k$$

- Properties
  - The Dirichlet distribution over K variables is confined to a K-1 dimensional simplex.
  - Expectations:



$$\begin{aligned} \mathbb{E}[\mu_k] &= \frac{\alpha_k}{\alpha_0} \\ \mathrm{var}[\mu_k] &= \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \\ \mathrm{cov}[\mu_j \mu_k] &= -\frac{\alpha_j \alpha_k}{\alpha_0^2(\alpha_0 + 1)} \end{aligned}$$

Dirichlet Distribution: Visualization



$\alpha_k = 10^{-1}$ $\alpha_k = 10^{0}$ $\alpha_k = 10^{1}$

Bayesian Multinomial

- Posterior distribution over the parameters $\{\mu_k\}$

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu}) p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

  - Comparison with the definition gives us the normalization factor

$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) &= \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m}) \\ &= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)\cdots\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \end{aligned}$$

⇒ We can interpret the parameters α_k of the Dirichlet prior as an effective number of observations of $x_k = 1$.

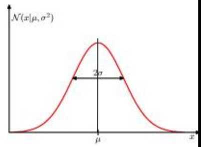
Topics of This Lecture

- Probability Distributions
  - Bayesian Estimation Reloaded
- Binary Variables
  - Bernoulli distribution
  - Binomial distribution
  - Beta distribution
- Multinomial Variables
  - Multinomial distribution
  - Dirichlet distribution
- Continuous Variables
  - Gaussian distribution
  - Gamma distribution
  - Student's t distribution
  - Exponential Family

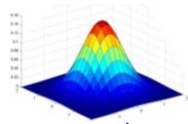
The Gaussian Distribution

- One-dimensional case
  - Mean μ
  - Variance σ^2

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



- Multi-dimensional case
  - Mean $\boldsymbol{\mu}$
  - Covariance $\boldsymbol{\Sigma}$

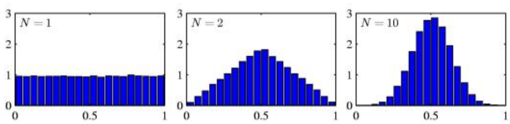


$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Image source: C.M. Bishop, 2006

Gaussian Distribution - Properties

- Central Limit Theorem
  - "The distribution of the sum of N i.i.d. random variables becomes increasingly Gaussian as N grows."
  - In practice, the convergence to a Gaussian can be very rapid.
  - This makes the Gaussian interesting for many applications.
- Example: N uniform [0,1] random variables.



Slide adapted from C. Bishop

Image source: C.M. Bishop, 2006

Gaussian Distribution - Properties

- Properties

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] = \boldsymbol{\Sigma}$$

- Limitations
  - Distribution is intrinsically unimodal, i.e. it is unable to provide a good approximation to multimodal distributions.
  - ⇒ We will see how to fix that with mixture distributions later…

Bayes' Theorem for Gaussian Variables

- Marginal and Conditional Gaussians
  - Suppose we are given a Gaussian prior $p(\mathbf{x})$ and a Gaussian conditional distribution $p(\mathbf{y}|\mathbf{x})$ (a linear Gaussian model)

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}\right)$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b},\mathbf{L}^{-1}\right)$$

  - From this, we can compute

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\},\boldsymbol{\Sigma})$$

  where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}$$

⇒ Closed-form solution for (Gaussian) marginal and posterior.

Slide adapted from C. Bishop

Maximum Likelihood for the Gaussian

- Maximum Likelihood
  - Given i.i.d. data $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)^T$, the log likelihood function is given by

$$\log p(\mathbf{X}|\mu,\boldsymbol{\Sigma}) = -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\mu)$$

- Sufficient statistics
  - The likelihood depends on the data set only through

$$\sum_{n=1}^{N}\mathbf{x}_n \qquad \sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}$$

  - Those are the sufficient statistics for the Gaussian distribution.

Slide adapted from C. Bishop

ML for the Gaussian

Advanced Machine Learning Winter'12

- Setting the derivative to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0$$

  - Solve to obtain

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n.$$

  - And similarly, but a bit more involved

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

ML for the Gaussian

Advanced Machine Learning Winter'12

- Comparison with true results
  - Under the true distribution, we obtain

$$\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] = \boldsymbol{\mu}$$
$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] = \frac{N-1}{N}\boldsymbol{\Sigma}.$$

  ⇒ The ML estimate for the covariance is biased and underestimates the true covariance!

  - Therefore define the following unbiased estimator

$$\widetilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

Bayesian Inference for the Gaussian

Advanced Machine Learning Winter'12

- Let's begin with a simple example
  - Consider a single Gaussian random variable x.
  - Assume σ^2 is known and the task is to infer the mean μ.
  - Given i.i.d. data $\mathbf{X} = (x_1, \ldots, x_N)^T$, the likelihood function for μ is given by

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

  - The likelihood function has a Gaussian shape as a function of μ.

  ⇒ The conjugate prior for this case is again a Gaussian.

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

Bayesian Inference for the Gaussian

Advanced Machine Learning Winter'12

- Combined with a Gaussian prior over μ

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

  - This results in the posterior

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$$

  - Completing the square over μ, we can derive that

$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$

where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

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

Visualization of the Results

Advanced Machine Learning Winter'12

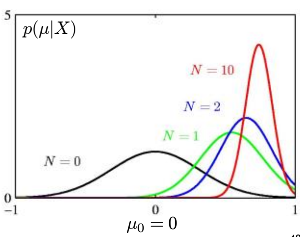
- Bayes estimate:

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

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

- Behavior for large N

	$N = 0$	$N \rightarrow \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_0^2	0



Bayesian Inference for the Gaussian

Advanced Machine Learning Winter'12

- More complex case
  - Now assume μ is known and the precision λ shall be inferred.
  - The likelihood function for $\lambda = 1/\sigma^2$ is given by

$$p(\mathbf{X}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

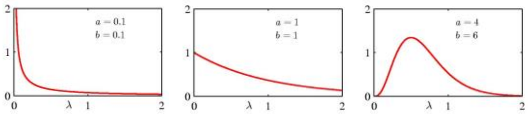
  - This has the shape of a Gamma function of λ.

The Gamma Distribution

- **Gamma distribution**
  - Product of a power of λ and the exponential of a linear function of λ.

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

- **Properties**
  - Finite integral if $a>0$ and the distribution itself is finite if $a \geq 1$.
  - Moments $\quad \mathbb{E}[\lambda] = \frac{a}{b} \qquad \mathrm{var}[\lambda] = \frac{a}{b^2}$
  - Visualization



Bayesian Inference for the Gaussian

- **Bayesian estimation**
  - Combine a Gamma prior $\mathrm{Gam}(\lambda|a_0, b_0)$ with the likelihood function to obtain

$$p(\lambda|\mathbf{X}) \propto \lambda^{a_0-1}\lambda^{N/2} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

  - We recognize this again as a Gamma function $\mathrm{Gam}(\lambda|a_N, b_N)$ with

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma^2_{\mathrm{ML}}.$$

Bayesian Inference for the Gaussian

- **Even more complex case**
  - Assume that both μ and λ are unknown
  - The joint likelihood function is given by

$$p(\mathbf{X}|\mu,\lambda) = \prod_{n=1}^{N}\left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\}$$

$$\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}.$$

$\Rightarrow$ Need a prior with the same functional dependence on μ and λ.

The Gaussian-Gamma Distribution

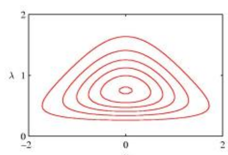
- **Gaussian-Gamma distribution**

$$p(\mu,\lambda) = \mathcal{N}(\mu|\mu_0,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b)$$

$$\propto \underbrace{\exp\left\{-\frac{\beta\lambda}{2}(\mu-\mu_0)^2\right\}}\ \underbrace{\lambda^{a-1}\exp\{-b\lambda\}}$$

  - Quadratic in μ.
  - Linear in λ.

- **Visualization**



Bayesian Inference for the Gaussian

- **Multivariate conjugate priors**
  - μ unknown, Λ known: $\quad p(\mu)$ Gaussian.
  - Λ unknown, μ known: $\quad p(\Lambda)$ Wishart,

$$\mathcal{W}(\mathbf{\Lambda}|\mathbf{W},\nu) = B|\mathbf{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\mathbf{\Lambda})\right).$$

  - Λ and μ unknown: $\quad p(\mu,\Lambda)$ Gaussian-Wishart,

$$p(\mu,\mathbf{\Lambda}|\mu_0,\beta,\mathbf{W},\nu) = \mathcal{N}(\mu||\mu_0,(\beta\mathbf{\Lambda})^{-1})\,\mathcal{W}(\mathbf{\Lambda}|\mathbf{W},\nu)$$

Student's t-Distribution

- **Gaussian estimation**
  - The conjugate prior for the precision of a Gaussian is a Gamma distribution.
  - Suppose we have a univariate Gaussian $\mathcal{N}(x|\mu,\tau^{-1})$ together with a Gamma prior $\mathrm{Gam}(\tau|a,b)$.
  - By integrating out the precision, obtain the marginal distribution

$$p(x|\mu,a,b) = \int_0^\infty \mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|a,b)\mathrm{d}\tau$$

$$= \int_0^\infty \mathcal{N}\left(x|\mu,(\eta\lambda)^{-1}\right)\mathrm{Gam}(\eta|\nu/2,\nu/2)\mathrm{d}\eta$$

  - This corresponds to an infinite mixture of Gaussians having the same mean, but different precision.

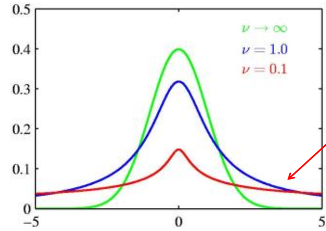
Student's t-Distribution

- **Student's t-Distribution**
  - We reparametrize the infinite mixture of Gaussians to get

$$\mathrm{St}(x|\mu,\lambda,\nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2}$$

- **Parameters**
  - "Precision" $\lambda = a/b$
  - "Degrees of freedom" $\nu = 2a$.

Student's t-Distribution: Visualization



Longer-tailed distribution!

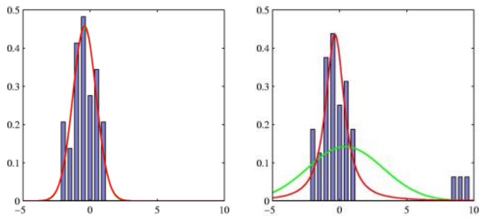
⇒ More robust to outliers…

- **Behavior**

	$\nu = 1$	$\nu \to \infty$
$\mathrm{St}(x\|\mu,\lambda,\nu)$	Cauchy	$\mathcal{N}(x\|\mu,\lambda^{-1})$

Student's t-Distribution

- Robustness to outliers: Gaussian vs t-distribution.



⇒ The t-distribution is much less sensitive to outliers, can be used for robust regression.

⇒ Downside: ML solution for t-distribution requires EM algorithm.

Student's t-Distribution: Multivariate Case

- Multivariate case in D dimensions

$$\begin{aligned}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) &= \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta \\ &= \frac{\Gamma(D/2+\nu/2)}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}}\left[1+\frac{\Delta^2}{\nu}\right]^{-D/2-\nu/2}\end{aligned}$$

where $\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})$ is the Mahalanobis distance.

- Properties

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu}, & &\text{if } \nu > 1 \\ \mathrm{cov}[\mathbf{x}] &= \frac{\nu}{(\nu-2)}\boldsymbol{\Lambda}^{-1}, & &\text{if } \nu > 2 \\ \mathrm{mode}[\mathbf{x}] &= \boldsymbol{\mu}\end{aligned}$$

References and Further Reading

- Probability distributions and their properties are described in Chapter 2 of Bishop's book.

Christopher M. Bishop
Pattern Recognition and Machine Learning
Springer, 2006