An Introduction to Probabilistic modeling

Oliver Stegle and Karsten Borgwardt

Machine Learning and
Computational Biology Research Group,
Max Planck Institute for Biological Cybernetics and
Max Planck Institute for Developmental Biology, Tübingen

Why probabilistic modeling?

- Inferences from data are intrinsically uncertain.
- Probability theory: model uncertainty instead of ignoring it!
- Applications: Machine learning, Data Mining, Pattern Recognition, etc.
- Goal of this part of the course
  - Overview on probabilistic modeling
  - Key concepts
  - Focus on Applications in Bioinformatics

Why probabilistic modeling?

- Inferences from data are intrinsically uncertain.
- Probability theory: model uncertainty instead of ignoring it!
- Applications: Machine learning, Data Mining, Pattern Recognition, etc.
- Goal of this part of the course
  - Overview on probabilistic modeling
  - Key concepts
  - Focus on Applications in Bioinformatics

Why probabilistic modeling?

- Inferences from data are intrinsically uncertain.
- Probability theory: model uncertainty instead of ignoring it!
- Applications: Machine learning, Data Mining, Pattern Recognition, etc.
- Goal of this part of the course
  - Overview on probabilistic modeling
  - Key concepts
  - Focus on Applications in Bioinformatics

Further reading, useful material

- Christopher M. Bishop: Pattern Recognition and Machine learning.
  - Good background, covers most of the course material and much more!
  - Substantial parts of this tutorial borrow figures and ideas from this book.
- David J.C. MacKay: Information Theory, Learning and Inference
  - Very worth while reading, not quite the same quality of overlap with the lecture synopsis.
  - Freely available online.

Lecture overview

1. An Introduction to probabilistic modeling
2. Applications: linear models, hypothesis testing
3. An introduction to Gaussian processes
4. Applications: time series, model comparison
5. Applications: continued

Outline

Outline

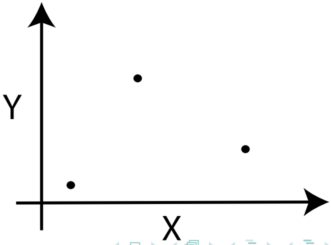
Key concepts
Data

- Let $\mathcal{D}$ denote a dataset, consisting of N datapoints
  $\mathcal{D} = \{\underbrace{\mathbf{x}_n}_{\text{Inputs}}, \underbrace{y_n}_{\text{Outputs}}\}_{n=1}^{N}$.
- Typical (this course)
  - $\mathbf{x} = \{x_1, \ldots, x_D\}$ multivariate, spanning D features for each observation (nodes in a graph, etc.).
  - y univariate (fitness, expression level etc.).
- Notation:
  - Scalars are printed as y.
  - Vectors are printed in bold: $\mathbf{x}$.
  - Matrices are printed in capital bold: $\mathbf{\Sigma}$.

Key concepts
Data

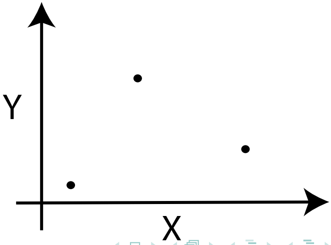
- Let $\mathcal{D}$ denote a dataset, consisting of N datapoints $\mathcal{D} = \{ \underbrace{\mathbf{x}_n}_{\text{Inputs}}, \underbrace{y_n}_{\text{Outputs}} \}_{n=1}^{N}$.
- Typical (this course)
  - $\mathbf{x} = \{x_1, \dots, x_D\}$ multivariate, spanning D features for each observation (nodes in a graph, etc.).
  - y univariate (fitness, expression level etc.).

- Notation:
  - Scalars are printed as y.
  - Vectors are printed in bold: $\mathbf{x}$.
  - Matrices are printed in capital bold: $\mathbf{\Sigma}$.



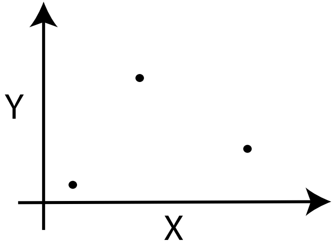
Key concepts
Data

- Let $\mathcal{D}$ denote a dataset, consisting of N datapoints $\mathcal{D} = \{\underbrace{\mathbf{x}_n}_{\text{Inputs}}, \underbrace{y_n}_{\text{Outputs}}\}_{n=1}^{N}$.
- Typical (this course)
  - $\mathbf{x} = \{x_1, \ldots, x_D\}$ multivariate, spanning D features for each observation (nodes in a graph, etc.).
  - y univariate (fitness, expression level etc.).
- Notation:
  - Scalars are printed as y.
  - Vectors are printed in bold: $\mathbf{x}$.
  - Matrices are printed in capital bold: $\mathbf{\Sigma}$.



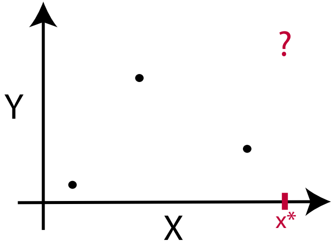
Key concepts
Predictions

- Observed dataset $\mathcal{D} = \{\underbrace{\mathbf{x}_n}_{\text{Inputs}}, \underbrace{y_n}_{\text{Outputs}}\}_{n=1}^{N}$.
- Given $\mathcal{D}$, what can we say about $y^{\star}$ at an unseen test input $\mathbf{x}^{\star}$?



Key concepts
Predictions

- Observed dataset $\mathcal{D} = \{\underbrace{\mathbf{x}_n}_{\text{Inputs}}, \underbrace{y_n}_{\text{Outputs}}\}_{n=1}^{N}$.
- Given $\mathcal{D}$, what can we say about $y^\star$ at an unseen test input $\mathbf{x}^\star$?

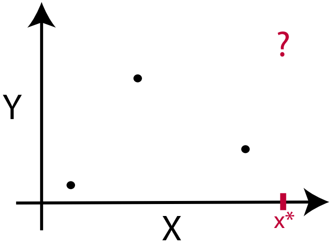


Key concepts
Model

- Observed dataset $\mathcal{D} = \{\underbrace{\mathbf{x}_n}_{\text{Inputs}}, \underbrace{y_n}_{\text{Outputs}}\}_{n=1}^{N}$.
- Given $\mathcal{D}$, what can we say about $y^\star$ at an unseen test input $\mathbf{x}^\star$?
- To make predictions we need to make assumptions.
- A model $\mathcal{H}$ encodes these assumptions and often depends on some parameters $\boldsymbol{\theta}$.

- Curve fitting: the model relates $\mathbf{x}$ to y,

$$y = f(x \,|\, \boldsymbol{\theta}) = \underbrace{\theta_0 + \theta_1 \cdot x}_{\text{example, a linear model}}$$

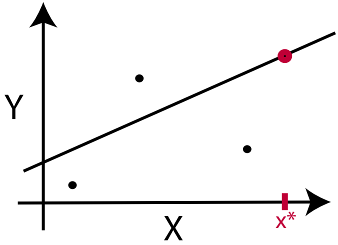


Key concepts
Model

- Observed dataset $\mathcal{D} = \{ \underbrace{\mathbf{x}_n}_{\text{Inputs}}, \underbrace{y_n}_{\text{Outputs}} \}_{n=1}^{N}$.
- Given $\mathcal{D}$, what can we say about $y^\star$ at an unseen test input $\mathbf{x}^\star$?
- To make predictions we need to make assumptions.
- A model $\mathcal{H}$ encodes these assumptions and often depends on some parameters $\boldsymbol{\theta}$.

- Curve fitting: the model relates $\mathbf{x}$ to y,

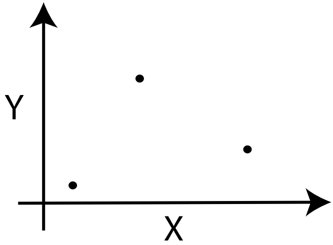
$$
\begin{aligned}
y &= f(x \,|\, \boldsymbol{\theta}) \\
&= \underbrace{\theta_0 + \theta_1 \cdot x}_{\text{example, a linear model}}
\end{aligned}
$$



Key concepts
Uncertainty

- Virtually in all steps there is uncertainty
  - Measurement uncertainty ($\mathcal{D}$)
  - Parameter uncertainty ($\boldsymbol{\theta}$)
  - Uncertainty regarding the correct model ($\mathcal{H}$)

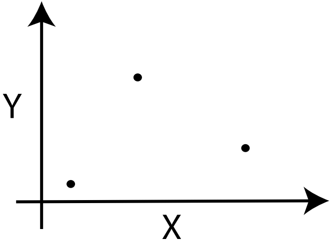
- Uncertainty can occur in both inputs and outputs.
- How to represent uncertainty?



Key concepts
Uncertainty

- Virtually in all steps there is uncertainty
  - Measurement uncertainty ($\mathcal{D}$)
  - Parameter uncertainty ($\boldsymbol{\theta}$)
  - Uncertainty regarding the correct model ($\mathcal{H}$)

- Uncertainty can occur in both inputs and outputs.
- How to represent uncertainty?

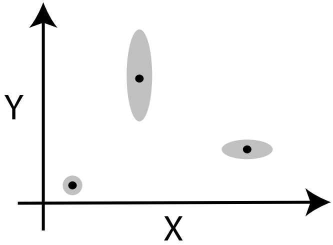


Key concepts
Uncertainty

- Virtually in all steps there is uncertainty
  - Measurement uncertainty ($\mathcal{D}$)
  - Parameter uncertainty ($\boldsymbol{\theta}$)
  - Uncertainty regarding the correct model ($\mathcal{H}$)

Measurement uncertainty

- Uncertainty can occur in both inputs and outputs.
- How to represent uncertainty?



Outline

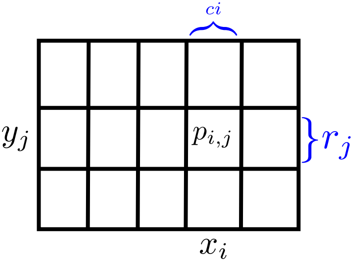
Probabilities

- Let X be a random variable, defined over a set $\mathcal{X}$ or measurable space.
- $P(X = x)$ denotes the probability that X takes value x, short $p(x)$.
  - Probabilities are positive, $P(X = x) \geq 0$
  - Probabilities sum to one

$$\int_{x \in \mathcal{X}} p(x) dx = 1 \qquad \sum_{x \in \mathcal{X}} p(x) = 1$$

- Special case: no uncertainty $p(x) = \delta(x - \hat{x})$.

Probability Theory



Joint Probability

$$P(X = x_i, Y = y_j) = \frac{n_{i,j}}{N}$$

Marginal Probability

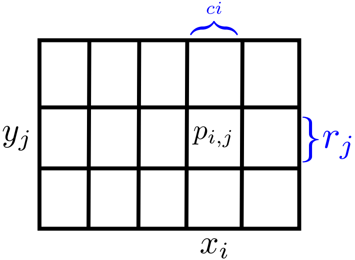
$$P(X = x_i) = \frac{c_i}{N}$$

Conditional Probability

$$P(Y = y_j \mid X = x_i) = \frac{n_{i,j}}{c_i}$$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Probability Theory



Marginal Probability

$$P(X = x_i) = \frac{c_i}{N}$$

Conditional Probability

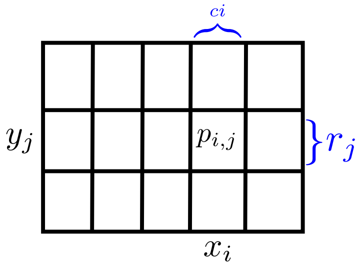
$$P(Y = y_j \mid X = x_i) = \frac{n_{i,j}}{c_i}$$

Product Rule

$$P(X = x_i, Y = y_j) = \frac{n_{i,j}}{N} = \frac{n_{i,j}}{c_i} \cdot \frac{c_i}{N}$$
$$= P(Y = y_j \mid X = x_i) P(X = x_i)$$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Probability Theory



Sum Rule

$$P(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{i,j}$$

$$= \sum_j P(X = x_i, Y = y_j)$$

Product Rule

$$P(X = x_i, Y = y_j) = \frac{n_{i,j}}{N} = \frac{n_{i,j}}{c_i} \cdot \frac{c_i}{N}$$

$$= P(Y = y_j \mid X = x_i) P(X = x_i)$$

(C.M. Bishop, Pattern Recognition and Machine Learning)

The Rules of Probability

Sum & Product Rule

Sum Rule $\quad p(x) = \sum_y p(x, y)$

Product Rule $\quad p(x, y) = p(y \mid x)p(x)$

The Rules of Probability

Bayes Theorem

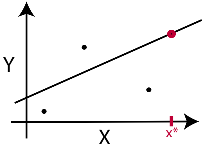
- Using the product rule we obtain

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$

$$p(x) = \sum_y p(x \mid y)p(y)$$

Bayesian probability calculus

- Bayes rule is the basis for inference and learning.
- Assume we have a model with parameters $\boldsymbol{\theta}$, e.g.



$$y = \theta_0 + \theta_1 \cdot x$$

- Goal: learn parameters $\boldsymbol{\theta}$ given Data $\mathcal{D}$.

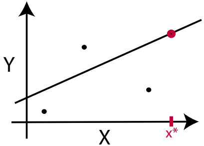
$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \boldsymbol{\theta}) \; p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

- Posterior
- Likelihood
- Prior

Bayesian probability calculus

- Bayes rule is the basis for inference and learning.
- Assume we have a model with parameters $\boldsymbol{\theta}$, e.g.

$$y = \theta_0 + \theta_1 \cdot x$$



- Goal: learn parameters $\boldsymbol{\theta}$ given Data $\mathcal{D}$.

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \boldsymbol{\theta}) \; p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

posterior $\propto$ likelihood $\cdot$ prior

- Posterior
- Likelihood
- Prior

Information and Entropy

- Information is the reduction of uncertainty.
- Entropy $H(X)$ is the quantitative description of uncertainty
  - $H(X) = 0$: certainty about X.
  - $H(X)$ maximal if all possibilities are equal probable.
  - Uncertainty and information are additive.
- These conditions are fulfilled by the entropy function:

$$H(X) = -\sum_{x \in \mathcal{X}} P(X = x) \log P(X = x)$$

Information and Entropy

- Information is the reduction of uncertainty.
- Entropy $H(X)$ is the quantitative description of uncertainty
  - $H(X) = 0$: certainty about X.
  - $H(X)$ maximal if all possibilities are equal probable.
  - Uncertainty and information are additive.
- These conditions are fulfilled by the entropy function:

$$H(X) = -\sum_{x \in \mathcal{X}} P(X = x) \log P(X = x)$$

Definitions related to entropy and information

- Entropy is the average surprise

$$H(X) = \sum_{x \in \mathcal{X}} P(X = x) \underbrace{\left(-\log P(X = x)\right)}_{\text{surprise}}$$

- Conditional entropy

$$H(X \mid Y) = - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(X = x, Y = y) \log P(X = x \mid Y = y)$$

- Mutual information

$$I(X : Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$
$$H(X) + H(Y) - H(X, Y)$$

- Independence of X and Y, $p(x, y) = p(x)p(y)$.

Definitions related to entropy and information

- Entropy is the average surprise

$$H(X) = \sum_{x \in \mathcal{X}} P(X = x) \underbrace{(-\log P(X = x))}_{\text{surprise}}$$

- Conditional entropy

$$H(X \mid Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(X = x, Y = y) \log P(X = x \mid Y = y)$$

- Mutual information

$$I(X : Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$
$$H(X) + H(Y) - H(X, Y)$$

- Independence of X and Y, $p(x, y) = p(x)p(y)$.

Definitions related to entropy and information

- Entropy is the average surprise

$$H(X) = \sum_{x \in \mathcal{X}} P(X = x) \underbrace{(-\log P(X = x))}_{\text{surprise}}$$

- Conditional entropy

$$H(X \mid Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(X = x, Y = y) \log P(X = x \mid Y = y)$$

- Mutual information

$$I(X : Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$
$$H(X) + H(Y) - H(X, Y)$$

- Independence of X and Y, $p(x, y) = p(x)p(y)$.

Definitions related to entropy and information

- Entropy is the average surprise

$$H(X) = \sum_{x \in \mathcal{X}} P(X = x) \underbrace{(-\log P(X = x))}_{\text{surprise}}$$

- Conditional entropy

$$H(X \mid Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(X = x, Y = y) \log P(X = x \mid Y = y)$$

- Mutual information

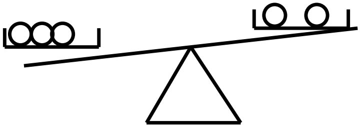
$$I(X : Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$
$$H(X) + H(Y) - H(X, Y)$$

- Independence of X and Y, $p(x, y) = p(x)p(y)$.

Entropy in action

The optimal weighting problem

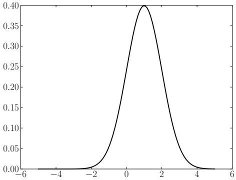
- Given 12 balls, all equal except for one that is lighter or heavier.
- What is the ideal weighting strategy and how many weightings are needed to identify the odd ball?



Probability distributions

- Gaussian

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



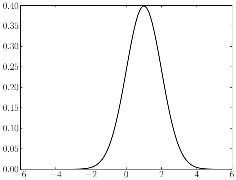
- Multivariate Gaussian

$$p(x \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$
$$= \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

Probability distributions

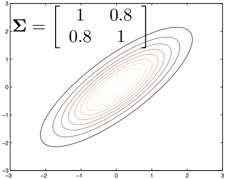
- Gaussian

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



- Multivariate Gaussian

$$p(x \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$= \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$



Probability distributions

continued...

- Bernoulli

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x}$$

- Gamma

$$p(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

Probability distributions

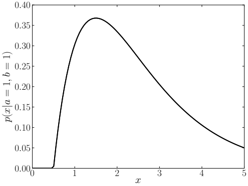
continued...

- Bernoulli

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x}$$

- Gamma

$$p(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$



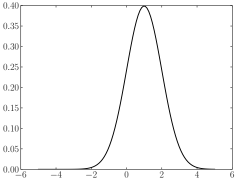
Probability distributions

The Gaussian revisited

- Gaussian PDF

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

- Positive: $\mathcal{N}\left(x \,\middle|\, \mu, \sigma^2\right) > 0$
- Normalized: $\int_{-\infty}^{+\infty} \mathcal{N}\left(x \,|\, \mu, \sigma\right) \mathrm{d}x = 1$ (check)
- Expectation: $< x > = \int_{-\infty}^{+\infty} \mathcal{N}\left(x \,\middle|\, \mu, \sigma^2\right) x \mathrm{d}x = \mu$
- Variance: $\mathrm{Var}[x] = < x^2 > - < x >^2$ $= \mu^2 + \sigma^2 - \mu^2 = \sigma^2$

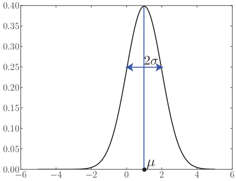


Probability distributions
The Gaussian revisited

- Gaussian PDF

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

- Positive: $\mathcal{N}\left(x \,\middle|\, \mu, \sigma^2\right) > 0$
- Normalized: $\int_{-\infty}^{+\infty} \mathcal{N}\left(x \,\middle|\, \mu, \sigma\right) \mathrm{d}x = 1$ (check)
- Expectation:
  $< x >= \int_{-\infty}^{+\infty} \mathcal{N}\left(x \,\middle|\, \mu, \sigma^2\right) x \mathrm{d}x = \mu$
- Variance: $\text{Var}[x] =< x^2 > - < x >^2$
  $= \mu^2 + \sigma^2 - \mu^2 = \sigma^2$



Outline

Inference for the Gaussian

Ingredients

- Data

$$\mathcal{D} = \{x_1, \dots, x_N\}$$

- Model $\mathcal{H}_{Gauss}$ – Gaussian PDF

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

$$\boldsymbol{\theta} = \{\mu, \sigma^2\}$$

- Likelihood

$$p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \,\middle|\, \mu, \sigma^2\right)$$

Inference for the Gaussian
Ingredients

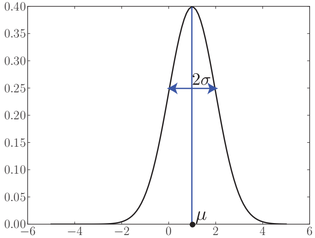
- Data

$$\mathcal{D} = \{x_1, \dots, x_N\}$$

- Model $\mathcal{H}_{Gauss}$ – Gaussian PDF

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

$$\boldsymbol{\theta} = \{\mu, \sigma^2\}$$



- Likelihood

$$p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \,\middle|\, \mu, \sigma^2\right)$$

Inference for the Gaussian
Ingredients

- Data

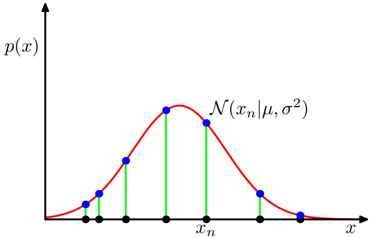
$$\mathcal{D} = \{x_1, \ldots, x_N\}$$

- Model $\mathcal{H}_{\mathsf{Gauss}}$ – Gaussian PDF

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

- Likelihood

$$p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \,\middle|\, \mu, \sigma^2\right)$$



(C.M. Bishop, Pattern Recognition and Machine Learning)

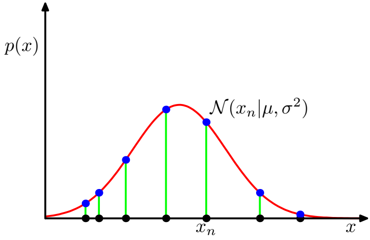
Inference for the Gaussian
Maximum likelihood

- Likelihood

$$p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \,|\, \mu, \sigma^2\right)$$

- Maximum likelihood

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D} \,|\, \boldsymbol{\theta})$$



(C.M. Bishop, Pattern Recognition and Machine Learning)

Inference for the Gaussian

Maximum likelihood

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D} \mid \boldsymbol{\theta}) \qquad = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2}$$

Inference for the Gaussian

Maximum likelihood

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ln p(\mathcal{D} \,|\, \boldsymbol{\theta}) \qquad = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ln \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n - \mu)^2}$$

Inference for the Gaussian

Maximum likelihood

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ln p(\mathcal{D} \mid \boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[-\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right]$$

Inference for the Gaussian

Maximum likelihood

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ln p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[-\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right]$$

$$\hat{\mu} : \frac{\mathrm{d}}{\mu} \ln p(\mathcal{D} \,|\, \mu) = 0 \qquad \hat{\sigma}^2 : \frac{\mathrm{d}}{\sigma^2} \ln p(\mathcal{D} \,|\, \sigma^2) = 0$$

Inference for the Gaussian

Maximum likelihood

Inference for the Gaussian

Maximum likelihood

- Maximum likelihood solutions

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

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

  Equivalent to common mean and variance estimators (almost).

- Maximum likelihood ignores parameter uncertainty
  - Think of the ML solution for a single observed datapoint x_1

$$\mu_{\mathsf{ML1}} = x_1$$

$$\sigma^2_{\mathsf{ML1}} = (x_1 - \mu_{ML1})^2 = 0$$

- How about Bayesian inference?

Inference for the Gaussian

Maximum likelihood

- Maximum likelihood solutions

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

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

  Equivalent to common mean and variance estimators (almost).
- Maximum likelihood ignores parameter uncertainty
  - Think of the ML solution for a single observed datapoint x_1

$$\mu_{\mathsf{ML1}} = x_1$$

$$\sigma^2_{\mathsf{ML1}} = (x_1 - \mu_{ML1})^2 = 0$$

- How about Bayesian inference?

Inference for the Gaussian

Maximum likelihood

- Maximum likelihood solutions

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

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

  Equivalent to common mean and variance estimators (almost).
- Maximum likelihood ignores parameter uncertainty
  - Think of the ML solution for a single observed datapoint x_1

$$\mu_{\mathsf{ML}1} = x_1$$
$$\sigma^2_{\mathsf{ML}1} = (x_1 - \mu_{ML1})^2 = 0$$

- How about Bayesian inference?

Bayesian Inference for the Gaussian

Ingredients

- Data

$$\mathcal{D} = \{x_1, \ldots, x_N\}$$

- Model $\mathcal{H}_{Gauss}$ – Gaussian PDF

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

$$\boldsymbol{\theta} = \{\mu\}$$

  - For simplicity: assume variance σ^2 is known.

- Likelihood

$$p(\mathcal{D} \mid \mu) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right)$$

Bayesian Inference for the Gaussian

Ingredients

- Data

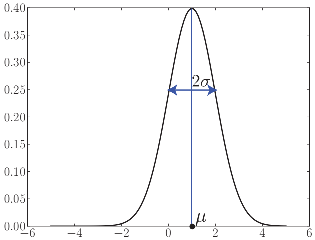
$$\mathcal{D} = \{x_1, \dots, x_N\}$$

- Model $\mathcal{H}_{Gauss}$ – Gaussian PDF

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

$$\boldsymbol{\theta} = \{\mu\}$$

  - For simplicity: assume variance σ^2 is known.



- Likelihood

$$p(\mathcal{D} \mid \mu) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right)$$

Bayesian Inference for the Gaussian
Ingredients

- Data

$$\mathcal{D} = \{x_1, \dots, x_N\}$$

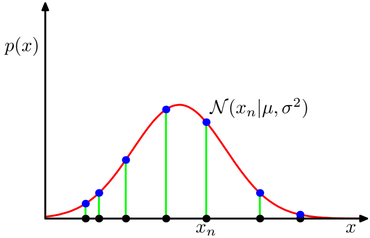
- Model $\mathcal{H}_{\mathsf{Gauss}}$ – Gaussian PDF

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

  - For simplicity: assume variance σ^2 is known.

- Likelihood

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



(C.M. Bishop, Pattern Recognition and Machine Learning)

Bayesian Inference for the Gaussian

Bayes rule

- Combine likelihood with a Gaussian prior over μ

$$p(\mu) = \mathcal{N}\left(\mu \,\middle|\, m_0, s_0^2\right)$$

- The posterior is proportional to

$$p(\mu \,|\, \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \,|\, \mu, \sigma^2) p(\mu)$$

Bayesian Inference for the Gaussian

$$p(\mu \mid \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \mid \mu) p(\mu)$$

$$= \left[\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n - \mu)^2} \right] \frac{1}{\sqrt{2\pi s_0^2}} e^{-\frac{1}{2s_0^2}(\mu - m_0)^2}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}^{N} \frac{1}{\sqrt{2\pi s_0^2}}}_{C1} \exp\left[-\frac{1}{2s_0^2}(\mu^2 - 2\mu m_0 + m_0^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (\mu^2 - 2\mu x_n + x_n^2) \right]$$

$$= C2 \exp\left[-\frac{1}{2} \underbrace{\left(\frac{1}{s_0^2} + \frac{N}{\sigma^2} \right)}_{1/\hat{\sigma}} \left(\mu^2 - 2\mu \underbrace{\hat{\sigma}\left(\frac{1}{s_0^2} m_0 + \frac{1}{\sigma^2} \sum_{n=1}^{N} x_n \right)}_{\hat{\mu}} \right) + C3 \right]$$

- Posterior parameters follow as the new coefficients.
- Note: All the constants we dropped on the way yield the model evidence: $p(\mu \mid \mathcal{D}, \sigma^2) = \frac{p(\mathcal{D} \mid \mu) p(\mu)}{Z}$

Bayesian Inference for the Gaussian

$$p(\mu \,|\, \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \,|\, \mu) p(\mu)$$

$$= \left[\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n - \mu)^2} \right] \frac{1}{\sqrt{2\pi s_0^2}} e^{-\frac{1}{2s_0^2}(\mu - m_0)^2}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}^{N} \frac{1}{\sqrt{2\pi s_0^2}}}_{C1} \exp\left[-\frac{1}{2s_0^2}(\mu^2 - 2\mu m_0 + m_0^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (\mu^2 - 2\mu x_n + x_n^2) \right]$$

$$= C2 \exp\left[-\frac{1}{2} \underbrace{\left(\frac{1}{s_0^2} + \frac{N}{\sigma^2} \right)}_{1/\hat{\sigma}} \left(\mu^2 - 2\mu \underbrace{\hat{\sigma}\left(\frac{1}{s_0^2} m_0 + \frac{1}{\sigma^2} \sum_{n=1}^{N} x_n\right)}_{\hat{\mu}} \right) + C3 \right]$$

- Posterior parameters follow as the new coefficients.
- Note: All the constants we dropped on the way yield the model evidence: $p(\mu \,|\, \mathcal{D}, \sigma^2) = \frac{p(\mathcal{D} \,|\, \mu) p(\mu)}{Z}$

Bayesian Inference for the Gaussian

$$p(\mu \mid \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \mid \mu) p(\mu)$$

$$= \left[\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n - \mu)^2} \right] \frac{1}{\sqrt{2\pi s_0^2}} e^{-\frac{1}{2s_0^2}(\mu - m_0)^2}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}^{N} \frac{1}{\sqrt{2\pi s_0^2}}}_{C1} \exp\left[-\frac{1}{2s_0^2}(\mu^2 - 2\mu m_0 + m_0^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (\mu^2 - 2\mu x_n + x_n^2) \right]$$

$$= C2 \exp\left[-\frac{1}{2} \underbrace{\left(\frac{1}{s_0^2} + \frac{N}{\sigma^2} \right)}_{1/\hat{\sigma}} \left(\mu^2 - 2\mu \underbrace{\hat{\sigma}\left(\frac{1}{s_0^2} m_0 + \frac{1}{\sigma^2} \sum_{n=1}^{N} x_n \right)}_{\hat{\mu}} \right) + C3 \right]$$

- Posterior parameters follow as the new coefficients.
- Note: All the constants we dropped on the way yield the model evidence: $p(\mu \mid \mathcal{D}, \sigma^2) = \frac{p(\mathcal{D} \mid \mu) p(\mu)}{Z}$

Bayesian Inference for the Gaussian

- Posterior of the mean: $p(\mu \mid \mathcal{D}, \sigma^2) \propto \mathcal{N}(\mu \mid \hat{\mu}, \hat{\sigma})$, after some rewriting

$$\hat{\mu} = \frac{\sigma^2}{N s_0^2 + \sigma^2} m_0 + \frac{N s_0^2}{N s_0^2 + \sigma^2} \mu_{\mathsf{ML}}, \quad \mu_{\mathsf{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

$$\frac{1}{\hat{\sigma}^2} = \frac{1}{s_0^2} + \frac{N}{\sigma^2}$$

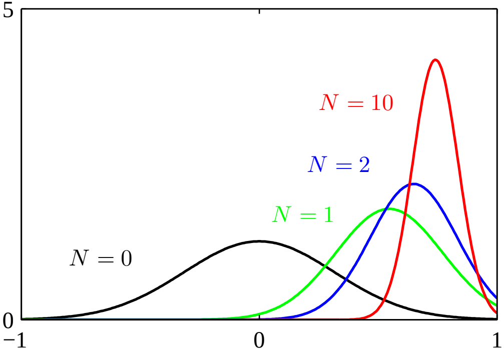
- Limiting cases for no and infinite amount of data

	$N = 0$	$N \to \infty$
$\hat{\mu}$	m_0	μ_{ML}
$\hat{\sigma}^2$	s_0^2	0

Bayesian Inference for the Gaussian

Examples

- Posterior $p(\mu \,|\, \mathcal{D}, \sigma^2)$ for increasing data sizes.



(C.M. Bishop, Pattern Recognition and Machine Learning)

Conjugate priors

- It is not chance that the posterior

$$p(\mu \,|\, \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \,|\, \mu, \sigma^2) p(\mu)$$

is tractable in closed form for the Gaussian.

Conjugate prior

$p(\theta)$ is a conjugate prior for a particular likelihood $p(\mathcal{D} \,|\, \theta)$ if the posterior is of the same functional form than the prior.

Conjugate priors

- It is not chance that the posterior

$$p(\mu \,|\, \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \,|\, \mu, \sigma^2) p(\mu)$$

is tractable in closed form for the Gaussian.

Conjugate prior

$p(\theta)$ is a conjugate prior for a particular likelihood $p(\mathcal{D} \,|\, \theta)$ if the posterior is of the same functional form than the prior.

Conjugate priors
Exponential family distributions

- A large class of probability distributions are part of the exponential family (all in this course) and can be written as:

$$p(\mathbf{x}\,|\,\boldsymbol{\theta}) = h(\mathbf{x})g(\boldsymbol{\theta})\exp\{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\}$$

- For example for the Gaussian:

$$\begin{aligned} p(x\,|\,\mu,\sigma^2) &= \frac{1}{2\pi\sigma^2}\exp\{-\frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2)\} \\ &= h(x)g(\boldsymbol{\theta})exp\{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\} \end{aligned}$$

with $\boldsymbol{\theta} = \left(\begin{array}{c} \mu/\sigma^2 \\ -1/2\sigma^2 \end{array} \right)$, $h(x) = \dfrac{1}{\sqrt{2\pi}}$

$\mathbf{u}(x) = \left(\begin{array}{c} x \\ x^2 \end{array} \right)$, $g(\boldsymbol{\theta}) = (-2\theta_2)^{1/2}\exp\left(\dfrac{\theta_1^2}{4\theta_2}\right)$

Conjugate priors

Exponential family distributions

Conjugacy and exponential family distributions

- For all members of the exponential family it is possible to construct a conjugate prior.
  - Intuition: The exponential form ensures that we can construct a prior that keeps its functional form.

- Conjugate priors for the Gaussian $\mathcal{N}\left(x \,\middle|\, \mu, \sigma^2\right)$
  - $p(\mu) = \mathcal{N}\left(\mu \,\middle|\, m_0, s_0^2\right)$
  - $p(\frac{1}{\sigma^2}) = \Gamma(\frac{1}{\sigma^2}, a_0, b_0).$

Conjugate priors
Exponential family distributions

Conjugacy and exponential family distributions

- For all members of the exponential family it is possible to construct a conjugate prior.
  - Intuition: The exponential form ensures that we can construct a prior that keeps its functional form.

- Conjugate priors for the Gaussian $\mathcal{N}\left(x \,\middle|\, \mu, \sigma^2\right)$
  - $p(\mu) = \mathcal{N}\left(\mu \,\middle|\, m_0, s_0^2\right)$
  - $p(\frac{1}{\sigma^2}) = \Gamma(\frac{1}{\sigma^2}, a_0, b_0)$.

Bayesian Inference for the Gaussian

Sequential learning

- Bayes rule naturally leads itself to sequential learning
- Assume one by one multiple datasets become available: $\mathcal{D}_1, \ldots, \mathcal{D}_S$

$$p_1(\boldsymbol{\theta}) \propto p(\mathcal{D}_1 \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})$$
$$p_2(\boldsymbol{\theta}) \propto p(\mathcal{D}_2 \,|\, \boldsymbol{\theta}) p_1(\boldsymbol{\theta})$$
$$\ldots$$

- Note: Assuming the datasets are independent, sequential updates and a single learning step yield the same answer.

Outline

Summary

- Probability theory: the language of uncertainty.
- Key rules of probability: sum rule, product rule.
- Bayes rules formes the fundamentals of learning. (posterior $\propto$ likelihood $\cdot$ prior).
- The entropy quantifies uncertainty.
- Parameter learning using maximum likelihood.
- Bayesian inference for the Gaussian.