Foundations of Machine Learning
African Masters in Machine Intelligence

Logistic Regression

Marc Deisenroth

Quantum Leap Africa
African Institute for Mathematical Sciences, Rwanda

Department of Computing
Imperial College London

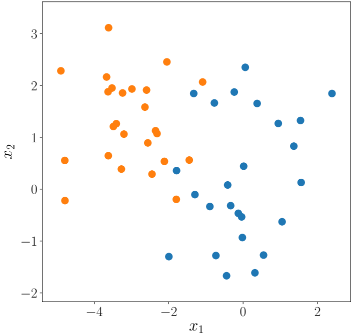
@mpd37
mdeisenroth@aimsammi.org

November 5, 2018

Learning Material

- Pattern Recognition and Machine Learning, Chapter 4 (Bishop, 2006)
- Machine Learning: A Probabilistic Perspective, Chapter 8 (Murphy, 2012)

Binary Classification



- Supervised learning setting with inputs $\boldsymbol{x}_n \in \mathbb{R}^D$ and binary targets $y_n \in \{0, 1\}$ belonging to classes $\mathcal{C}_1, \mathcal{C}_2$.
- Objective: Find a decision boundary/surface that separates the two classes as well as possible

Class Posteriors

- Binary classification problem with two classes $\mathcal{C}_1, \mathcal{C}_2$.

Class Posteriors

- Binary classification problem with two classes $\mathcal{C}_1, \mathcal{C}_2$.
- Posterior class probability $p(y = 1|\boldsymbol{x}) = p(\mathcal{C}_1|\boldsymbol{x})$:

$$p(\mathcal{C}_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x})},$$

$$p(\boldsymbol{x}) = p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)$$

Class Posteriors

- Binary classification problem with two classes $\mathcal{C}_1, \mathcal{C}_2$.
- Posterior class probability $p(y=1|\boldsymbol{x}) = p(\mathcal{C}_1|\boldsymbol{x})$:

$$p(\mathcal{C}_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x})},$$

$$p(\boldsymbol{x}) = p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)$$

- Define the log-ratio of the posteriors (log-odds)

$$a := \log \frac{p(\mathcal{C}_1|\boldsymbol{x})}{p(\mathcal{C}_2|\boldsymbol{x})} = \log \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

Class Posteriors

- Binary classification problem with two classes $\mathcal{C}_1, \mathcal{C}_2$.
- Posterior class probability $p(y = 1|\boldsymbol{x}) = p(\mathcal{C}_1|\boldsymbol{x})$:

$$p(\mathcal{C}_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x})},$$

$$p(\boldsymbol{x}) = p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)$$

- Define the log-ratio of the posteriors (log-odds)

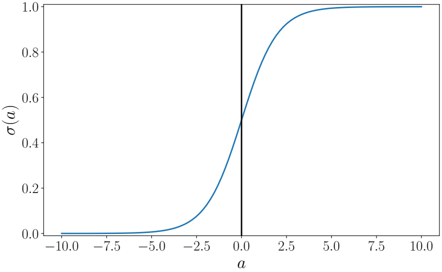
$$a := \log \frac{p(\mathcal{C}_1|\boldsymbol{x})}{p(\mathcal{C}_2|\boldsymbol{x})} = \log \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

- Then

$$\underbrace{\sigma(a) := \frac{1}{1 + \exp(-a)}}_{\text{logistic sigmoid}} =?$$

▸▸ **Discuss with your neighbors**

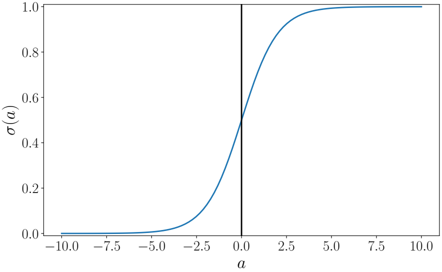
Logistic Sigmoid



$$a := \log \frac{p(\mathcal{C}_1|\boldsymbol{x})}{p(\mathcal{C}_2|\boldsymbol{x})} = \log \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

$$\sigma(a) := \frac{1}{1+\exp(-a)} = p(\mathcal{C}_1|\boldsymbol{x})$$ **Logistic sigmoid**

Logistic Sigmoid



$$a := \log \frac{p(\mathcal{C}_1|\boldsymbol{x})}{p(\mathcal{C}_2|\boldsymbol{x})} = \log \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

$$\sigma(a) := \frac{1}{1+\exp(-a)} = p(\mathcal{C}_1|\boldsymbol{x})$$ **Logistic sigmoid**

- Assign the label for $\mathcal{C}_1$ to $\boldsymbol{x}$ if $\sigma(a) = p(\mathcal{C}_1|\boldsymbol{x}) = p(y=1|\boldsymbol{x}) \geqslant 0.5$

Generalization to the Multiclass Setting

- Assume we are given K classes. Then

$$p(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_{j=1}^{K} p(\boldsymbol{x}|\mathcal{C}_j)p(\mathcal{C}_j)}$$

  is the generalization of the logistic sigmoid to K classes.
- **Softmax function, Boltzmann distribution, normalized exponential**

Implicit Modeling Assumptions

- Assume Gaussian class conditionals

$$p(\boldsymbol{x}|\mathcal{C}_k) = \mathcal{N}\big(\boldsymbol{x}\,|\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}\big)$$

where the covariance matrix $\boldsymbol{\Sigma}$ is shared across all K classes.

Implicit Modeling Assumptions

- Assume Gaussian class conditionals

$$p(\boldsymbol{x}|\mathcal{C}_k) = \mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\big)$$

  where the covariance matrix $\boldsymbol{\Sigma}$ is shared across all K classes.

- For $K = 2$ we get (Bishop, 2006)

$$p(\mathcal{C}_1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x} + \theta_0)\,,$$

$$\boldsymbol{\theta} := \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,, \quad \theta_0 := \frac{1}{2}\Big(\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\Big) + \log\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

Implicit Modeling Assumptions

- Assume Gaussian class conditionals

$$p(\boldsymbol{x}|\mathcal{C}_k) = \mathcal{N}\big(\boldsymbol{x}\,|\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}\big)$$

  where the covariance matrix $\boldsymbol{\Sigma}$ is shared across all K classes.

- For $K = 2$ we get (Bishop, 2006)

$$p(\mathcal{C}_1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x} + \theta_0)\,,$$

$$\boldsymbol{\theta} := \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,, \quad \theta_0 := \frac{1}{2}\Big(\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\Big) + \log \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

  ▸▸ Argument of the sigmoid is linear in $\boldsymbol{x}$

Implicit Modeling Assumptions

- Assume Gaussian class conditionals

$$p(\boldsymbol{x}|\mathcal{C}_k) = \mathcal{N}\big(\boldsymbol{x}\,|\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}\big)$$

where the covariance matrix $\boldsymbol{\Sigma}$ is shared across all K classes.

- For $K = 2$ we get (Bishop, 2006)

$$p(\mathcal{C}_1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x} + \theta_0)\,,$$

$$\boldsymbol{\theta} := \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,, \quad \theta_0 := \frac{1}{2}\Big(\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\Big) + \log \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

▶▶ Argument of the sigmoid is linear in $\boldsymbol{x}$

▶▶ Decision boundary is a surface along which the posterior class probabilities $p(\mathcal{C}_k|\boldsymbol{x})$ are constant

▶▶ **Decision boundary is a linear function of $\boldsymbol{x}$**

Implicit Modeling Assumptions

- Assume Gaussian class conditionals

$$p(\boldsymbol{x}|\mathcal{C}_k) = \mathcal{N}\big(\boldsymbol{x}\,|\,\boldsymbol{\mu}_k,\,\boldsymbol{\Sigma}\big)$$

where the covariance matrix $\boldsymbol{\Sigma}$ is shared across all K classes.

- For $K = 2$ we get (Bishop, 2006)

$$p(\mathcal{C}_1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x} + \theta_0)\,,$$

$$\boldsymbol{\theta} := \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,, \quad \theta_0 := \frac{1}{2}\Big(\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\Big) + \log \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

▸▸ Argument of the sigmoid is linear in $\boldsymbol{x}$

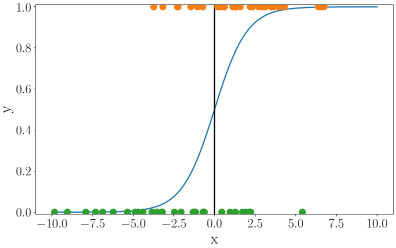
▸▸ Decision boundary is a surface along which the posterior class probabilities $p(\mathcal{C}_k|\boldsymbol{x})$ are constant

▸▸ **Decision boundary is a linear function of $\boldsymbol{x}$**

- If covariances are not shared: Quadratic decision boundaries

Model Specification (Logistic Regression)

likelihood



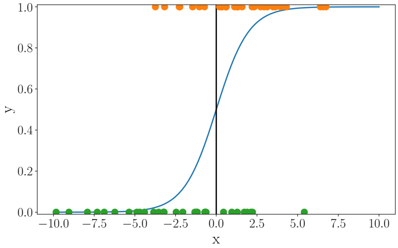
Model Specification (Logistic Regression)

- Bernoulli likelihood

$$y \in \{0, 1\}$$

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(y|\mu(\boldsymbol{x}))\,,$$

$$\mu(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x})$$



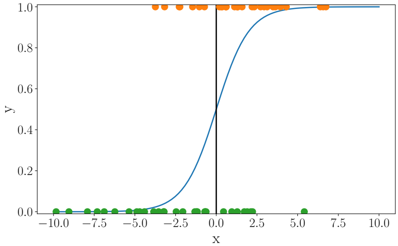
Model Specification (Logistic Regression)

- Bernoulli likelihood

$$y \in \{0, 1\}$$

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(y|\mu(\boldsymbol{x}))\,,$$

$$\mu(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x})$$



- Label y depends on input location $\boldsymbol{x}$, i.e., $\mu(\boldsymbol{x})$ needs to be a function of $\boldsymbol{x}$

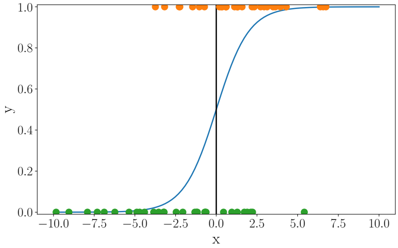
Model Specification (Logistic Regression)

- Bernoulli likelihood

$y \in \{0, 1\}$

$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(y|\mu(\boldsymbol{x}))\,,$

$\mu(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x})$



- Label y depends on input location $\boldsymbol{x}$, i.e., $\mu(\boldsymbol{x})$ needs to be a function of $\boldsymbol{x}$
- Idea: Linear model $\boldsymbol{\theta}^\top \boldsymbol{x}$ (as in linear regression)

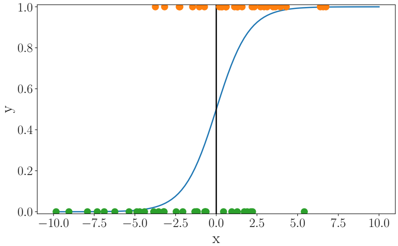
Model Specification (Logistic Regression)

- Bernoulli likelihood

$y \in \{0, 1\}$

$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(y|\mu(\boldsymbol{x}))\,,$

$\mu(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x})$



- Label y depends on input location $\boldsymbol{x}$, i.e., $\mu(\boldsymbol{x})$ needs to be a function of $\boldsymbol{x}$
- Idea: Linear model $\boldsymbol{\theta}^\top \boldsymbol{x}$ (as in linear regression)
- Ensure $0 \leqslant \mu(\boldsymbol{x}) \leqslant 1$

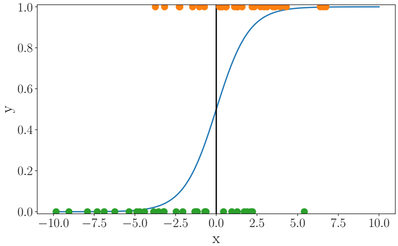
Model Specification (Logistic Regression)

- Bernoulli likelihood

$y \in \{0, 1\}$

$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \text{Ber}(y|\mu(\boldsymbol{x}))\,,$

$\mu(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x})$



- Label y depends on input location $\boldsymbol{x}$, i.e., $\mu(\boldsymbol{x})$ needs to be a function of $\boldsymbol{x}$
- Idea: Linear model $\boldsymbol{\theta}^\top \boldsymbol{x}$ (as in linear regression)
- Ensure $0 \leqslant \mu(\boldsymbol{x}) \leqslant 1$
- Squash the linear combination through a function that guarantees this: $\mu(\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x})$

$$\Longrightarrow p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \text{Ber}(y|\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}))$$

Model Fitting

- Estimate model parameters $\boldsymbol{\theta}$ (MLE or MAP)

Model Fitting

- Estimate model parameters $\boldsymbol{\theta}$ (MLE or MAP)
- Likelihood (training data $\boldsymbol{X}, \boldsymbol{y}$):

Model Fitting

- Estimate model parameters $\boldsymbol{\theta}$ (MLE or MAP)
- Likelihood (training data $\boldsymbol{X}, \boldsymbol{y}$):

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) &= \prod_{n=1}^{N} \mathrm{Ber}(y_n|\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}_n)) = \prod_{n=1}^{N} (\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}_n))^{y_n} (1 - \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}_n))^{1-y_n} \\
&= \prod_{n=1}^{N} \mu_n^{y_n} (1 - \mu_n)^{1-y_n} \\
\mu_n &:= \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}_n)
\end{aligned}
$$

- **Negative log likelihood (cross-entropy):**

Model Fitting

- Estimate model parameters $\boldsymbol{\theta}$ (MLE or MAP)
- Likelihood (training data $\boldsymbol{X}, \boldsymbol{y}$):

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \text{Ber}(y_n|\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}_n)) = \prod_{n=1}^{N} (\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}_n))^{y_n} (1 - \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}_n))^{1-y_n}$$

$$= \prod_{n=1}^{N} \mu_n^{y_n} (1 - \mu_n)^{1-y_n}$$

$$\mu_n := \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}_n)$$

- **Negative log likelihood (cross-entropy):**

$$NLL = -\sum_{n=1}^{N} y_n \log \mu_n + (1 - y_n) \log(1 - \mu_n)$$

Model Fitting (2)

- Derivative of sigmoid w.r.t. its argument:

$$\sigma(z_n) = \frac{1}{1 + \exp(-z_n)}$$
$$\implies \frac{\mathrm{d}\sigma(z_n)}{\mathrm{d}z_n} =$$

Model Fitting (2)

- Derivative of sigmoid w.r.t. its argument:

$$\sigma(z_n) = \frac{1}{1+\exp(-z_n)}$$
$$\implies \frac{\mathrm{d}\sigma(z_n)}{\mathrm{d}z_n} = \frac{\exp(-z_n)}{(1+\exp(-z_n))^2} = \sigma(z_n)(1-\sigma(z_n))$$

Model Fitting (2)

- Derivative of sigmoid w.r.t. its argument:

$$\sigma(z_n) = \frac{1}{1+\exp(-z_n)}$$
$$\implies \frac{\mathrm{d}\sigma(z_n)}{\mathrm{d}z_n} = \frac{\exp(-z_n)}{(1+\exp(-z_n))^2} = \sigma(z_n)(1-\sigma(z_n))$$

- Gradient of the negative log-likelihood:

$$\frac{\mathrm{d}NLL}{\mathrm{d}\boldsymbol{\theta}} = -\sum_{n=1}^{N}\left(y_n\frac{1}{\mu_n} - (1-y_n)\frac{1}{1-\mu_n}\right)\frac{\mathrm{d}\mu_n}{\mathrm{d}\boldsymbol{\theta}}$$
$$\frac{\mathrm{d}\mu_n}{\mathrm{d}\boldsymbol{\theta}} =$$

Model Fitting (2)

- Derivative of sigmoid w.r.t. its argument:

$$\sigma(z_n) = \frac{1}{1+\exp(-z_n)}$$

$$\implies \frac{\mathrm{d}\sigma(z_n)}{\mathrm{d}z_n} = \frac{\exp(-z_n)}{(1+\exp(-z_n))^2} = \sigma(z_n)(1-\sigma(z_n))$$

- Gradient of the negative log-likelihood:

$$\frac{\mathrm{d}NLL}{\mathrm{d}\boldsymbol{\theta}} = -\sum_{n=1}^{N}\left(y_n\frac{1}{\mu_n} - (1-y_n)\frac{1}{1-\mu_n}\right)\frac{\mathrm{d}\mu_n}{\mathrm{d}\boldsymbol{\theta}}$$

$$\frac{\mathrm{d}\mu_n}{\mathrm{d}\boldsymbol{\theta}} = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\sigma(\underbrace{\boldsymbol{\theta}^\top \boldsymbol{x}_n}_{z_n}) = \frac{\mathrm{d}\sigma(z_n)}{\mathrm{d}z_n}\frac{\mathrm{d}z_n}{\mathrm{d}\boldsymbol{\theta}} = \sigma(z_n)(1-\sigma(z_n))\boldsymbol{x}_n^\top$$

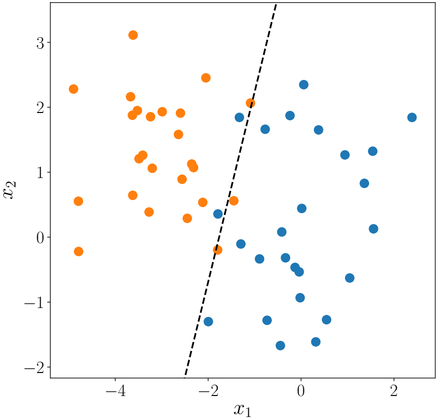
Model Fitting (3)

$$\frac{\mathrm{d}NLL}{\mathrm{d}\boldsymbol{\theta}} = (\boldsymbol{\mu} - \boldsymbol{y})^\top \boldsymbol{X}$$

$$\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]^\top$$

- No closed-form solution ⏩ Gradient descent methods
- Unique global optimum exists

Example

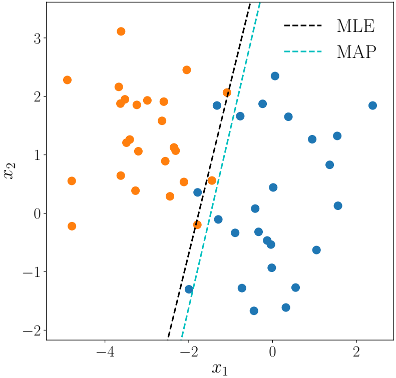


$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(\sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2))$$

Comments on Maximum Likelihood

- If the classes are linearly separable, the decision boundary is not unique and the likelihood will tend to infinity
- Overfitting is a again a problem when we work with features $\boldsymbol{\phi}(\boldsymbol{x})$ instead of $\boldsymbol{x}$
- Maximum a posteriori estimation can address these issues to some degree

MAP Estimation



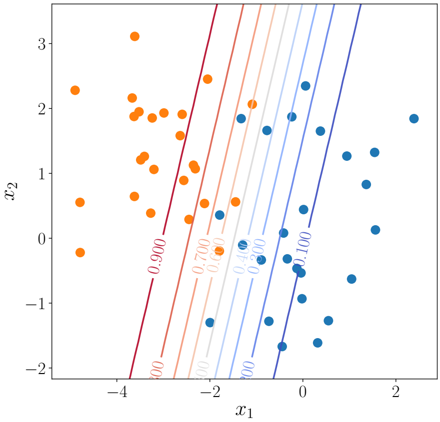
- Log-posterior:

$$\log p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \text{ const}$$

- No closed-form solution for $\boldsymbol{\theta}_{\text{MAP}}$

  ▶▶ Numerical maximization of the log-posterior

Predictive Labels



$$p(y = 1|\boldsymbol{x}, \boldsymbol{\theta}_{\text{MAP}}) = \text{Ber}(\sigma(\boldsymbol{x}^\top \boldsymbol{\theta}_{\text{MAP}}))$$

Bayesian Logistic Regression

Objective

For a given (i.i.d.) dataset $\mathcal{D} := \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ compute a posterior distribution on the parameters $\boldsymbol{\theta}$

Bayesian Logistic Regression

Objective

For a given (i.i.d.) dataset $\mathcal{D} := \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ compute a posterior distribution on the parameters $\boldsymbol{\theta}$

- Choose Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{0},\, \boldsymbol{S}_0\big)$
- Posterior (via Bayes' theorem):

Bayesian Logistic Regression

Objective

For a given (i.i.d.) dataset $\mathcal{D} := \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ compute a posterior distribution on the parameters $\boldsymbol{\theta}$

- Choose Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \,|\, \boldsymbol{0},\, \boldsymbol{S}_0)$
- Posterior (via Bayes' theorem):

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X})}{p(\boldsymbol{y}|\boldsymbol{X})} = \frac{\mathcal{N}(\boldsymbol{\theta} \,|\, \boldsymbol{0},\, \boldsymbol{S}_0) \prod_{n=1}^N \text{Ber}(\sigma(\boldsymbol{x}_n^\top \boldsymbol{\theta}))}{\int \mathcal{N}(\boldsymbol{\theta} \,|\, \boldsymbol{0},\, \boldsymbol{S}_0) \prod_{n=1}^N \text{Ber}(\sigma(\boldsymbol{x}_n^\top \boldsymbol{\theta})) \mathrm{d}\boldsymbol{\theta}}$$

Bayesian Logistic Regression

Objective

For a given (i.i.d.) dataset $\mathcal{D} := \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N)\}$ compute a posterior distribution on the parameters $\boldsymbol{\theta}$

- Choose Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{0},\, \boldsymbol{S}_0\big)$
- Posterior (via Bayes' theorem):

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X})}{p(\boldsymbol{y}|\boldsymbol{X})} = \frac{\mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{0},\, \boldsymbol{S}_0\big)\prod_{n=1}^{N} \mathrm{Ber}(\sigma(\boldsymbol{x}_n^\top \boldsymbol{\theta}))}{\int \mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{0},\, \boldsymbol{S}_0\big)\prod_{n=1}^{N} \mathrm{Ber}(\sigma(\boldsymbol{x}_n^\top \boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}}$$

- **No analytic solution**
  - ▶▶ Approximations necessary

Laplace Approximation

- Objective: Approximate an unknown distribution

$$p(\boldsymbol{x}) \propto \exp(-E(\boldsymbol{x})) =: \tilde{p}(\boldsymbol{x})$$

with a Gaussian distribution $q(\boldsymbol{x})$.

Laplace Approximation

- Objective: Approximate an unknown distribution

$$p(\boldsymbol{x}) \propto \exp(-E(\boldsymbol{x})) =: \tilde{p}(\boldsymbol{x})$$

  with a Gaussian distribution $q(\boldsymbol{x})$.
- Idea: Taylor-series expansion of $-\log \tilde{p}(\boldsymbol{x}) = E(\boldsymbol{x})$ around the mode $\boldsymbol{x}^*$ (MAP estimate)

Laplace Approximation

- Objective: Approximate an unknown distribution

$$p(\boldsymbol{x}) \propto \exp(-E(\boldsymbol{x})) =: \tilde{p}(\boldsymbol{x})$$

with a Gaussian distribution $q(\boldsymbol{x})$.

- Idea: Taylor-series expansion of $-\log \tilde{p}(\boldsymbol{x}) = E(\boldsymbol{x})$ around the mode $\boldsymbol{x}^*$ (MAP estimate)

$$-\log \tilde{p}(\boldsymbol{x}) \approx E(\boldsymbol{x}^*) + \boldsymbol{J}(\boldsymbol{x}^*)(\boldsymbol{x} - \boldsymbol{x}^*) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_*)^\top \boldsymbol{H}(\boldsymbol{x}_*)(\boldsymbol{x} - \boldsymbol{x}^*) ,$$

$$\boldsymbol{J}: \quad \text{Jacobian}, \qquad \boldsymbol{H}: \quad \text{Hessian}$$

Laplace Approximation

- Objective: Approximate an unknown distribution

$$p(\boldsymbol{x}) \propto \exp(-E(\boldsymbol{x})) =: \tilde{p}(\boldsymbol{x})$$

with a Gaussian distribution $q(\boldsymbol{x})$.

- Idea: Taylor-series expansion of $-\log \tilde{p}(\boldsymbol{x}) = E(\boldsymbol{x})$ around the mode $\boldsymbol{x}^*$ (MAP estimate)

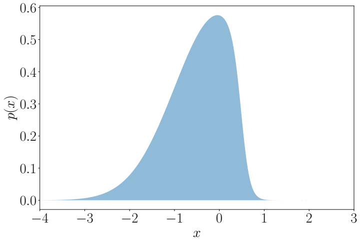
$$-\log \tilde{p}(\boldsymbol{x}) \approx E(\boldsymbol{x}^*) + \boldsymbol{J}(\boldsymbol{x}^*)(\boldsymbol{x} - \boldsymbol{x}^*) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_*)^\top \boldsymbol{H}(\boldsymbol{x}_*)(\boldsymbol{x} - \boldsymbol{x}^*)\,,$$

$$\boldsymbol{J}: \quad \text{Jacobian}, \qquad \boldsymbol{H}: \quad \text{Hessian}$$

- $J(\boldsymbol{x}^*) = \boldsymbol{0}^\top$ because $\boldsymbol{x}^*$ is a stationary point (mode) of $\log \tilde{p}$

$$\begin{aligned}\tilde{p}(\boldsymbol{x}) &\approx \exp(-E(\boldsymbol{x}^*)) \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_*)^\top \boldsymbol{H}(\boldsymbol{x}_*)(\boldsymbol{x} - \boldsymbol{x}^*)) \\ &\propto \mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{x}^*, \boldsymbol{H}^{-1}\big) =: q(\boldsymbol{x})\end{aligned}$$

Laplace Approximation: Example

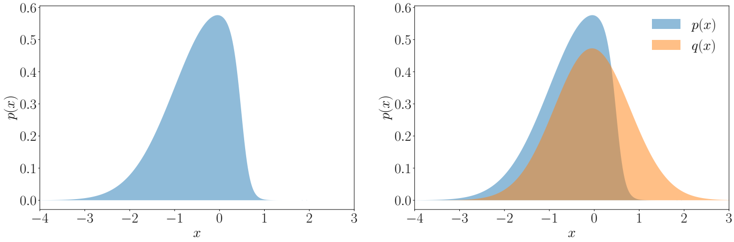


- Unnormalized distribution:

$$\tilde{p}(x) = \exp(-\tfrac{1}{2}x^2)\sigma(ax + b)$$

Discuss with your neighbors

Laplace Approximation: Example



- Unnormalized distribution:

$$\tilde{p}(x) = \exp(-\tfrac{1}{2}x^2)\sigma(ax+b)$$

$$q(x) = \mathcal{N}\left(x \,\middle|\, x^*, (1 + a^2\mu_*(1-\mu_*))^{-1}\right), \quad \mu_* := \sigma(ax_* + b)$$

Laplace Approximation: Properties

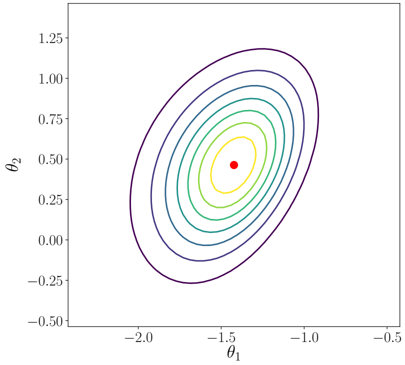
- Only need to know the unnormalized distribution $\tilde{p}$
- Finding the mode: numerical methods (optimization problem)
- Captures only local properties of the distribution
- Multimodal distributions: Approximation will be different depending on which mode we are in (not unique)

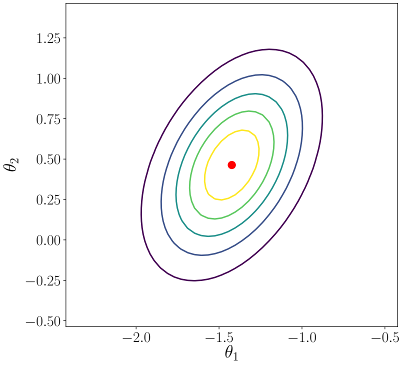
Laplace Approximation: Properties

- Only need to know the unnormalized distribution $\tilde{p}$
- Finding the mode: numerical methods (optimization problem)
- Captures only local properties of the distribution
- Multimodal distributions: Approximation will be different depending on which mode we are in (not unique)
- For large datasets, we would expect the posterior to converge to a Gaussian (central limit theorem)

  ▶▶ Laplace approximation should work well in this case

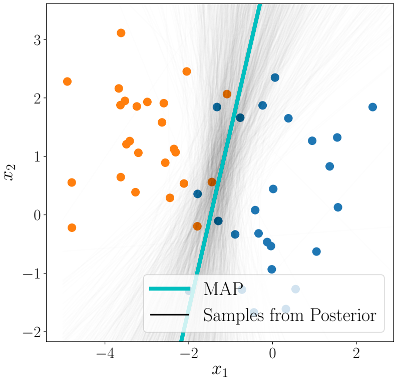
Posterior Approximation





- Left: true parameter posterior
- Right: Laplace approximation

Posterior Decision Boundary



- Parameter samples $\boldsymbol{\theta}_i$ drawn from Laplace approximation $q(\boldsymbol{\theta})$ of posterior $p(\boldsymbol{\theta}|\boldsymbol{X})$
- Decision boundary drawn for each $\boldsymbol{\theta}_i$

Predictions

Assume a Gaussian distribution $p(\boldsymbol{\theta}) = \mathcal{N}\big(\boldsymbol{\mu}, \boldsymbol{\Sigma}\big)$ on the parameters (e.g., Laplace approximation of the posterior). Then:

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{x}) &= \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&= \int \mathrm{Ber}(\sigma(\boldsymbol{\theta}^\top \boldsymbol{x})) \mathcal{N}\big(\boldsymbol{\theta} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}\big) \mathrm{d}\boldsymbol{\theta} \\
&= \mathbb{E}_{\boldsymbol{\theta}}[\mathrm{Ber}(\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}))]
\end{aligned}
$$

Predictions

Assume a Gaussian distribution $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on the parameters (e.g., Laplace approximation of the posterior). Then:

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{x}) &= \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&= \int \mathrm{Ber}(\sigma(\boldsymbol{\theta}^\top \boldsymbol{x})) \mathcal{N}(\boldsymbol{\theta} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\theta} \\
&= \mathbb{E}_{\boldsymbol{\theta}}[\mathrm{Ber}(\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}))]
\end{aligned}
$$

▶▶ **Integral intractable**

Predictions

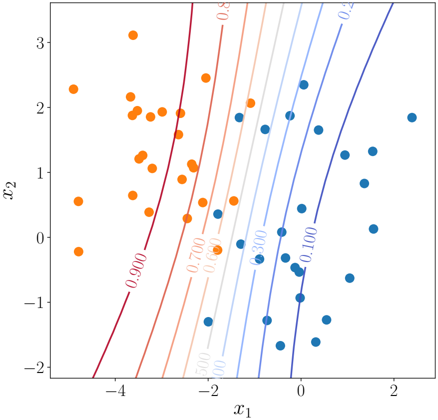
Assume a Gaussian distribution $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on the parameters (e.g., Laplace approximation of the posterior). Then:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x}) &= \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\ &= \int \mathrm{Ber}(\sigma(\boldsymbol{\theta}^\top \boldsymbol{x})) \mathcal{N}(\boldsymbol{\theta} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\theta} \\ &= \mathbb{E}_{\boldsymbol{\theta}}[\mathrm{Ber}(\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}))] \end{aligned}$$

▶▶ **Integral intractable**

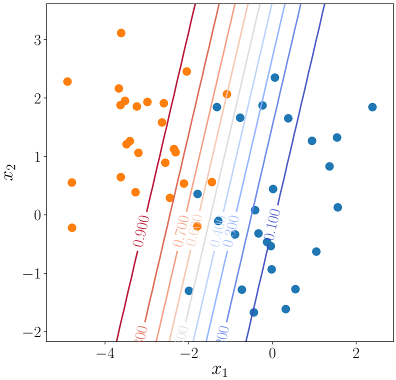
- "Plug-in approximation": use posterior mean (MAP estimate) $\mathbb{E}[\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}]$
- Monte Carlo estimate (sampling from $p(\boldsymbol{\theta})$ is easy)

Predictions (2)

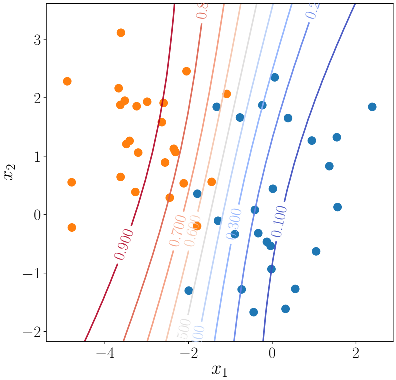


1. Samples from Laplace approximation of the posterior
2. Monte-Carlo estimate of label prediction

Comparison with MAP Predictions



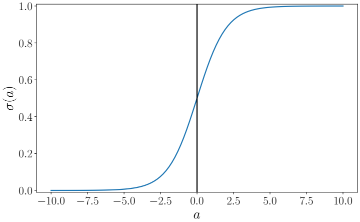
(a) MAP

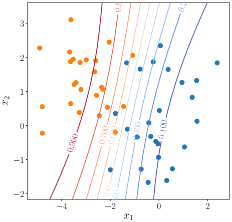


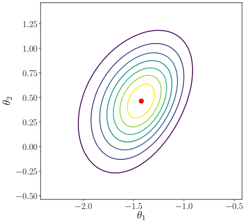
(b) Bayesian Logistic Regression

- Predictive labels

Summary







- Binary classification problems
- Linear model with non-Gaussian likelihood
- Implicit modeling assumptions
- Parameter estimation (MLE, MAP) no longer in closed form
- Bayesian logistic regression with Laplace approximation of the posterior

References I

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag, 2006.

[2] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA, USA, 2012.