Maximum likelihood estimation

Nuno Vasconcelos
UCSD
Bayesian decision theory

• recall that we have
  – Y – state of the world
  – X – observations
  – g(x) – decision function
  – L[g(x), y] – loss of predicting y with g(x)

• Bayes decision rule is the rule that minimizes the risk

\[
Risk = E_{X,Y}[L(X,Y)]
\]

• for the “0-1” loss

\[
L[g(x), y] = \begin{cases} 
1, & g(x) \neq y \\
0, & g(x) = y
\end{cases}
\]

• optimal decision rule is the maximum a-posteriori probability rule
MAP rule
• we have shown that it can be implemented in any of the three following ways
  – 1) \[ i^*(x) = \arg \max_i P_{Y|x}(i \mid x) \]
  – 2) \[ i^*(x) = \arg \max_i \left[ P_{X|Y}(x \mid i) P_Y(i) \right] \]
  – 3) \[ i^*(x) = \arg \max_i \left[ \log P_{X|Y}(x \mid i) + \log P_Y(i) \right] \]
• by introducing a “model” for the class-conditional distributions we can express this as a simple equation
  – e.g. for the multivariate Gaussian
    \[ P_{X|Y}(x \mid i) = \frac{1}{\sqrt{(2\pi)^d \left| \Sigma_i \right|}} \exp \left\{ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right\} \]
The Gaussian classifier

• the solution is

\[ i^*(x) = \arg \min_i [d_i(x, \mu_i) + \alpha_i] \]

with

\[ d_i(x, y) = (x - y)^T \Sigma_i^{-1} (x - y) \]

\[ \alpha_i = \log(2\pi)^d |\Sigma_i| - 2 \log P_Y(i) \]

• the optimal rule is to assign \( x \) to the closest class
• closest is measured with the Mahalanobis distance \( d_i(x, y) \)
• can be further simplified in special cases
Geometric interpretation

- for Gaussian classes, equal covariance $\sigma^2 I$

$$
\begin{align*}
    w &= \frac{\mu_i - \mu_j}{\sigma^2} \\
    x_0 &= \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j)
\end{align*}
$$
Geometric interpretation

- for Gaussian classes, equal but arbitrary covariance

\[ w = \Sigma^{-1}(\mu_i - \mu_j) \]

\[ x_0 = \frac{\mu_i + \mu_j}{2} - \frac{1}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j) \]
Bayesian decision theory

• advantages:
  – BDR is optimal and cannot be beaten
  – Bayes keeps you honest
  – models reflect causal interpretation of the problem, this is how we think
  – natural decomposition into “what we knew already” (prior) and “what data tells us” (CCD)
  – no need for heuristics to combine these two sources of info
  – BDR is, almost invariably, intuitive
  – Bayes rule, chain rule, and marginalization enable modularity, and scalability to very complicated models and problems

• problems:
  – BDR is optimal only insofar the models are correct.
Implementation

• we do have an optimal solution

\[ W = \Sigma^{-1}(\mu_i - \mu_j) \]
\[ x_0 = \frac{\mu_i + \mu_j}{2} - \frac{1}{\Sigma^{-1}(\mu_i - \mu_j)} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j) \]

• but in practice we do not know the values of the parameters \( \mu, \Sigma, P_Y(1) \)
  – we have to somehow estimate these values
  – this is OK, we can come up with an estimate from a training set
  – e.g. use the average value as an estimate for the mean

\[ w = \hat{\Sigma}^{-1}(\hat{\mu}_i - \hat{\mu}_j) \]
\[ x_0 = \frac{\hat{\mu}_i + \hat{\mu}_j}{2} - \frac{1}{\Sigma^{-1}(\hat{\mu}_i - \hat{\mu}_j)} \log \frac{\hat{P}_Y(i)}{\hat{P}_Y(j)} (\hat{\mu}_i - \hat{\mu}_j) \]
Important

• **warning:** at this point all optimality claims for the BDR cease to be valid!!

• the BDR is guaranteed to achieve the minimum loss when we use the true probabilities

• when we “plug in” the probability estimates, we could be implementing a classifier that is quite distant from the optimal
  - e.g. if the $P_{X|Y}(x|i)$ look like the example above
  - I could never approximate them well by parametric models (e.g. Gaussian)
Maximum likelihood

• this seems pretty serious
  – how should I get these probabilities then?
• we rely on the maximum likelihood (ML) principle
• this has three steps:
  – 1) we choose a parametric model for all probabilities
  – to make this clear we denote the vector of parameters by $\Theta$ and the class-conditional distributions by $P_{X|Y}(x \mid i; \Theta)$
    – note that this means that $\Theta$ is NOT a random variable (otherwise it would have to show up as subscript)
    – it is simply a parameter, and the probabilities are a function of this parameter
Maximum likelihood

• three steps:
  – 2) we assemble a collection of datasets
    \[ D^{(i)} = \{x_1^{(i)}, \ldots, x_n^{(i)}\} \]
    set of examples drawn independently from class i
  – 3) we select the parameters of class i to be the ones that maximize the probability of the data from that class

\[ \Theta_i = \arg \max_{\Theta} P_{X|Y}(D^{(i)} | i; \Theta) \]

\[ = \arg \max_{\Theta} \log P_{X|Y}(D^{(i)} | i; \Theta) \]

– like before, it does not really make any difference to maximize probabilities or their logs
Maximum likelihood

• since
  – each sample $\mathcal{D}^{(i)}$ is considered independently
  – parameter $\Theta_i$ estimated only from sample $\mathcal{D}^{(i)}$

• we simply have to repeat the procedure for all classes

• so, from now on we omit the class variable

$$
\Theta^* = \arg \max_{\Theta} P_X(D; \Theta)
$$

$$
= \arg \max_{\Theta} \log P_X(D; \Theta)
$$

• the function $P_X(D; \Theta)$ is called the likelihood of the parameter $\Theta$ with respect to the data

• or simply the likelihood function
Maximum likelihood

- note that the likelihood function is a function of the parameters $\Theta$
- it does not have the same shape as the density itself
- e.g. the likelihood function of a Gaussian is not bell-shaped
- the likelihood is defined only after we have a sample

$$P_X(d; \Theta) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left\{-\frac{(d - \mu)^2}{2\sigma^2}\right\}$$
Maximum likelihood

• given a sample, to obtain ML estimate we need to solve

\[ \Theta^* = \arg \max_{\Theta} P_X(D; \Theta) \]

• when \( \Theta \) is a scalar this is high-school calculus

we have a maximum when
  – first derivative is zero
  – second derivative is negative
The gradient

• in higher dimensions, the generalization of the derivative is the gradient
• the gradient of a function $f(w)$ at $z$ is

$$\nabla f(z) = \left( \frac{\partial f}{\partial w_0}(z), \ldots, \frac{\partial f}{\partial w_{n-1}}(z) \right)^T$$

• the gradient has a nice geometric interpretation
  – it points in the direction of maximum growth of the function
  – which makes it perpendicular to the contours where the function is constant
The gradient

• note that if $\nabla f = 0$
  – there is no direction of growth
  – also - $\nabla f = 0$, and there is no direction of decrease
  – we are either at a local minimum or maximum or “saddle” point

• conversely, at local min or max or saddle point
  – no direction of growth or decrease
  – $\nabla f = 0$

• this shows that we have a critical point if and only if $\nabla f = 0$

• to determine which type we need second order conditions
The Hessian

- the extension of the second-order derivative is the Hessian matrix

\[
\nabla^2 f(x) = \begin{bmatrix}
\frac{\partial^2 f}{\partial x_0^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_0 \partial x_{n-1}}(x) \\
\frac{\partial^2 f}{\partial x_{n-1} \partial x_0}(x) & \cdots & \frac{\partial^2 f}{\partial x_{n-1}^2}(x)
\end{bmatrix}
\]

- at each point \( x \), gives us the quadratic function

\[
x^t \nabla^2 f(x) x
\]

that best approximates \( f(x) \)
The Hessian

• this means that, when gradient is zero at $x$, we have
  – a maximum when function can be approximated by an “upwards-facing” quadratic
  – a minimum when function can be approximated by a “downwards-facing” quadratic
  – a saddle point otherwise
The Hessian

• for any matrix $M$, the function $\mathbf{x}^t M \mathbf{x}$ is
  — upwards facing quadratic when $M$ is negative definite
  — downwards facing quadratic when $M$ is positive definite
  — saddle otherwise

• hence, all that matters is the positive definiteness of the Hessian

• we have a maximum when Hessian is negative definite
Maximum likelihood

• in summary, given a sample, we need to solve

\[ \Theta^* = \arg \max_{\Theta} P_X(D; \Theta) \]

• the solutions are the parameters such that

\[ \nabla_{\Theta} P_X(D; \Theta) = 0 \]

\[ \theta^t \nabla_{\Theta}^2 P_X(D; \Theta) \theta \leq 0, \quad \forall \theta \in \mathbb{R}^n \]

• note that you always have to check the second-order condition!
Maximum likelihood

• let’s consider the Gaussian example

\[ f(T) = \frac{1}{\sigma_T \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{T - \bar{T}}{\sigma_T} \right)^2} \]

• given a sample \( \{T_1, \ldots, T_N\} \) of independent points
• the likelihood is

\[
L(T_1, T_2, \ldots, T_N | \bar{T}, \sigma_T) = L = \prod_{i=1}^{N} \left[ \frac{1}{\sigma_T \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{T_i - \bar{T}}{\sigma_T} \right)^2} \right] = \\
\frac{1}{(\sigma_T \sqrt{2\pi})^N} e^{-\frac{1}{2} \sum_{i=1}^{N} \left( \frac{T_i - \bar{T}}{\sigma_T} \right)^2}
\]
Maximum likelihood

• and the log-likelihood is

\[ \Lambda = \ln L = -\frac{N}{2} \ln(2\pi) - N \ln \sigma_T - \frac{1}{2} \sum_{i=1}^{N} \left( \frac{T_i - \bar{T}}{\sigma_T} \right)^2 \]

• the derivative with respect to the mean is zero when

\[ \frac{\partial(\Lambda)}{\partial \bar{T}} = \frac{1}{\sigma_T^2} \sum_{i=1}^{N} (T_i - \bar{T}) = 0 \]

• or

\[ \bar{T} = \frac{1}{N} \sum_{i=1}^{N} T_i \]

• note that this is just the sample mean
Maximum likelihood

• and the log-likelihood is

\[ \Lambda = \ln L = -\frac{N}{2} \ln(2\pi) - N \ln \sigma_T - \frac{1}{2} \sum_{i=1}^{N} \left( \frac{T_i - \bar{T}}{\sigma_T} \right)^2 \]

• the derivative with respect to the variance is zero when

\[ \frac{\partial (\Lambda)}{\partial \sigma_T} = -\frac{N}{\sigma_T} + \frac{1}{\sigma_T^3} \sum_{i=1}^{N} (T_i - \bar{T})^2 = 0 \]

• or

\[ \hat{\sigma}_T^2 = \frac{1}{N} \sum_{i=1}^{N} (T_i - \bar{T})^2 \]

• note that this is just the sample variance
Maximum likelihood

- example:
  - if sample is \{10,20,30,40,50\}

\[
\bar{T} = \frac{1}{N} \sum_{i=1}^{N} T_i
\]

\[
= \frac{10 + 20 + 30 + 40 + 50}{5}
\]

\[
= 30
\]

\[
\hat{\sigma}_T = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (T_i - \bar{T})^2}
\]

\[
= \sqrt{\frac{(10 - 30)^2 + (20 - 30)^2 + (30 - 30)^2 + (40 - 30)^2 + (50 - 30)^2}{5}}
\]

\[
= 14.1421
\]
Homework

• show that the Hessian is negative definite
  \[ \theta^T \nabla_\theta \nabla_\theta P_x(D; \theta) \theta \leq 0, \quad \forall \theta \in \mathbb{R}^n \]

• show that these formulas can be generalized to the vector case
  – \( D^{(i)} = \{x_1^{(i)}, \ldots, x_n^{(i)}\} \) set of examples from class \( i \)
  – the ML estimates are
  \[ \mu_i = \frac{1}{n} \sum_j x_j^{(i)} \]
  \[ \Sigma_i = \frac{1}{n} \sum_j (x_j^{(i)} - \mu_i)(x_j^{(i)} - \mu_i)^T \]

• note that the ML solution is usually intuitive
Estimators

- when we talk about estimators, it is important to keep in mind that
  - an estimate is a number
  - an estimator is a random variable

\[ \hat{\theta} = f(X_1, \ldots, X_n) \]

- an estimate is the value of the estimator for a given sample.
- if \( \mathcal{D} = \{x_1, \ldots, x_n\} \), when we say \( \hat{\mu} = \frac{1}{n} \sum x_j \)
  what we mean is \( \hat{\mu} = f(X_1, \ldots, X_n) \bigg|_{X_1=x_1, \ldots, X_n=x_n} \) with

\[ f(X_1, \ldots, X_n) = \frac{1}{n} \sum X_j \]

the \( X_i \) are random variables
Bias and variance

- we know how to produce estimators (by ML)
- how do we evaluate an estimator?
- Q₁: is the expected value equal to the true value?
- this is measured by the bias
  - if
    \[ \hat{θ} = f(X₁, \ldots, Xₙ) \]
  then
    \[ \text{Bias}(\hat{θ}) = E_{X₁, \ldots, Xₙ} \left[ f(X₁, \ldots, Xₙ) - θ \right] \]
  - an estimator that has bias will usually not converge to the perfect estimate \( θ \), no matter how large the sample is
  - e.g. if \( θ \) is negative and the estimator is
    \[ f(X₁, \ldots, Xₙ) = \frac{1}{n} \sum_j X_j^2 \]
  the bias is clearly non-zero
Bias and variance

• the estimators is said to be **biased**
  – this means that it is **not expressive** enough to approximate the true value arbitrarily well
  – this will be clearer when we talk about density estimation

• $Q_2$: assuming that the estimator converges to the true value, **how many sample points do we need?**
  – this can be measured by the **variance**

\[
\text{Var}(\hat{\theta}) = \left[ E_{X_1, \ldots, X_n} \left( f(X_1, \ldots, X_n) - E_{X_1, \ldots, X_n} [f(X_1, \ldots, X_n)] \right)^2 \right]
\]

– the variance **usually decreases** as one collects more training examples
Example

• ML estimator for the mean of a Gaussian $N(\mu, \sigma^2)$

$$
Bias(\hat{\mu}) = E_{X_1, \ldots, X_n}[\hat{\mu} - \mu] = E_{X_1, \ldots, X_n}[\hat{\mu}] - \mu
$$

$$
= E_{X_1, \ldots, X_n}\left[\frac{1}{n} \sum_i X_i\right] - \mu
$$

$$
= \frac{1}{n} \sum_i E_{X_1, \ldots, X_n}[X_i] - \mu
$$

$$
= \frac{1}{n} \sum_i E_{X_i}[X_i] - \mu
$$

$$
= \mu - \mu = 0
$$

• the estimator is unbiased
Example

- ML estimator for the mean of a Gaussian $N(\mu, \sigma^2)$

$$\text{Var}(\hat{\mu}) = E_{X_1, \ldots, X_n} \left\{ \left( \hat{\mu} - E_{X_1, \ldots, X_n}[\hat{\mu}] \right)^2 \right\} = E_{X_1, \ldots, X_n} \left\{ (\hat{\mu} - \mu)^2 \right\}$$

$$= E_{X_1, \ldots, X_n} \left\{ \left( \frac{1}{n} \sum_i X_i - \mu \right)^2 \right\}$$

$$= \frac{1}{n^2} E_{X_1, \ldots, X_n} \left\{ \left( \sum_i (X_i - \mu) \right)^2 \right\}$$

$$= \frac{1}{n^2} E_{X_1, \ldots, X_n} \left\{ \sum_{ij} (X_i - \mu)(X_j - \mu) \right\}$$
Example

• ML estimator for the mean of a Gaussian $N(\mu, \sigma^2)$

$$Var(\hat{\mu}) = \frac{1}{n^2} \sum_{ij} E_{X_i, X_j} [(X_i - \mu)(X_j - \mu)]$$

$$= \frac{1}{n^2} \sum_{ij} \sigma_{ij}$$

• and since $X_i, X_j$ are independent, $\sigma_{ij} = 0, \ \forall i \neq j$

$$Var(\hat{\mu}) = \frac{1}{n^2} \sum_i \sigma_i^2 = \frac{\sigma^2}{n}$$

• the variance goes to zero as $n$ increases!
Example

• in summary, for ML estimator for the mean of a Gaussian $N(\mu, \sigma^2)$

\[ E[\hat{\mu}] = \mu \quad Var(\hat{\mu}) = \frac{\sigma^2}{n} \]

• this means that if I have a large sample, the value of the estimate will be close to the true value with high probability
Example

• is this always true?
• ML estimator for the variance of a Gaussian $N(\mu, \sigma^2)$
  \[ \hat{\sigma}^2 = \frac{1}{n} \sum_i (X_i - \hat{\mu})^2 = \frac{1}{n} \sum_i \left( X_i^2 - 2X_i \hat{\mu} + \hat{\mu}^2 \right) \]

\[ = \frac{1}{n} \sum_i X_i^2 - \hat{\mu}^2 \]

• the expected value is
  \[ E_{X_1, \ldots, X_n} [\hat{\sigma}^2] = \frac{1}{n} \sum_i E_{X_1, \ldots, X_n} [X_i^2] - E_{X_1, \ldots, X_n} [\hat{\mu}^2] \]

\[ = \frac{1}{n} \sum_i E_{X_i} [X_i^2] - E_{X_1, \ldots, X_n} [\hat{\mu}^2] = E_X [X^2] - E_{X_1, \ldots, X_n} [\hat{\mu}^2] \]
Example

- using

\[
E_{X_1,\ldots,X_n}[\hat{\mu}^2] = E_{X_1,\ldots,X_n}\left[ \frac{1}{n^2} \sum_{ij} X_i X_j \right] = \frac{1}{n^2} \sum_{ij} E_{X_i,X_j}[X_i X_j]
\]

\[
= \frac{1}{n^2} \sum_i E_{X_i}[X_i^2] + \frac{1}{n^2} \sum_{i,j \neq i} E_{X_i,X_j}[X_i X_j]
\]

\[
= \frac{1}{n} E_X[X^2] + \frac{1}{n^2} \sum_i E_{X_i}[X_i] E_{X_j}[X_j]
\]

\[
= \frac{1}{n} E_X[X^2] + \frac{1}{n^2} \sum_i E_{X_i}[X_i] \sum_{j \neq i} E_{X_j}[X_j]
\]

\[
= \frac{1}{n} E_X[X^2] + \frac{1}{n^2} \sum_i E_{X_i}[X_i] (n-1) E_X[X]
\]

34
Example

• using

\[ E_{X_1,\ldots,X_n}[\hat{\mu}^2] = \frac{1}{n} E_X[X^2] + \frac{1}{n^2} \sum_i E_{X_i}[X_i](n-1)E_X[X] \]
\[ = \frac{1}{n} E_X[X^2] + \frac{(n-1)}{n} (E_X[X])^2 \]
\[ = \frac{1}{n} E_X[X^2] + \frac{(n-1)}{n} \mu^2 \]

• we get

\[ E_{X_1,\ldots,X_n}[\hat{\sigma}^2] = E_X[X^2] - E_{X_1,\ldots,X_n}[\hat{\mu}^2] \]
\[ = \frac{n-1}{n} E_X[X^2] - \frac{n-1}{n} \mu^2 = \left(1 - \frac{1}{n}\right)\sigma^2 \]
Example

• in summary

\[
E_{X_1, \ldots, X_n} \left[ \hat{\sigma}^2 \right] = \left( 1 - \frac{1}{n} \right) \sigma^2
\]

• the estimator is biased

• Q: do we care?
  – clearly

\[
\lim_{n \to \infty} E_{X_1, \ldots, X_n} \left[ \hat{\sigma}^2 \right] = \sigma^2
\]

  – so, for large samples it is (for all practical purposes) unbiased
  – what about small samples? the variance is likely to be large to start with, a little bit of bias is not going to make much difference
  – so, in practice, it is fine
Conclusion

• we really care about the conjunction of the two factors
  – working hard to decrease variance if bias is large is useless

\[ \hat{\theta} \]
\[ E[\theta] \]
\[ \theta \]
\[ n \]

– working hard to decrease bias if variance is large is useless

\[ E[\theta] = \theta \]
Any questions?